Constrained Estimation Using Penalization and MCMC*

A. Ronald Gallant Han Hong Michael P. Leung Jessie Li

November 14, 2020

Abstract. We study inference for parameters defined by either classical extremum estimators or Laplace-type estimators subject to general nonlinear constraints on the parameters. We show that running MCMC on the penalized version of the problem offers a computationally attractive alternative to solving the original constrained optimization problem. Bayesian credible intervals are asymptotically valid confidence intervals in a pointwise sense, providing exact asymptotic coverage for general functions of the parameters. We allow for nonadaptive and adaptive penalizations using the ℓ_p for $p \geqslant 1$ penalty functions. These methods are motivated by and include as special cases model selection and shrinkage methods such as the LASSO and its Bayesian and adaptive versions. A simulation study validates the theoretical results. We also provide an empirical application on estimating the joint density of U.S. real consumption and asset returns subject to Euler equation constraints in a CRRA asset pricing model.

Keywords: penalized estimation, MCMC, Laplace-type estimators, Bayesian LASSO
JEL Codes: C10, C11, C13, C15

*Department of Economics, Pennsylvania State University; Stanford University; University of Southern California; University of California, Santa Cruz. We thank the editor, the anonymous referees, Valentin Verdier, and participants in conferences and seminars for helpful comments. Han Hong acknowledges support by the National Science Foundation (SES 1164589) and SIEPR.

1 Introduction

The theoretical properties of constrained estimators when the objective function and constraints are smoothly differentiable and the parameters are uniquely identified over the unconstrained parameter space are already well understood (see e.g. Gallant, 1987). Furthermore, Hansen (2016) has shown the importance of applying shrinkage to nonlinearly constrained parameter spaces. However, in many settings, the objective function can be nonsmooth or nonconvex, or the parameters might be uniquely identified only over the constrained parameter space, in which case directly solving the constrained optimization problem through nonlinear programming methods remains computationally challenging in practice.

In this paper, we propose instead to adopt a Bayesian approach that defines penalized Laplace type estimators which are asymptotically equivalent to the original constrained estimators. Although the frequentist properties of unpenalized Laplace type estimators are well studied (e.g. Chernozhukov and Hong, 2003), the properties of penalized Laplace type estimators remain unknown. Similar to the unpenalized versions, penalized Laplace type estimators are typically defined as the mean or median of the quasi-posterior distribution of the parameters simulated using Markov Chain Monte Carlo (MCMC) methods. Our penalized Laplace type estimators include as a special case the Bayesian LASSO of Park and Casella (2008), who define their estimator using either the posterior mean or median in a Gaussian linear regression model with a Laplace prior. Furthermore, while much of the existing LASSO literature focuses on the Gaussian linear regression model, this paper considers general nonlinear and non-likelihood-based models, such as those in GMM, empirical likelihood, and minimum distance methods and allows for general nonlinear constraints for which traditional frequentist estimators are difficult to compute, since they require maximizing a possibly nonconvex or nonsmooth objective function.

We find that the penalized posterior mean and median are $\sqrt{n}$-consistent for the true parameters and achieve first order asymptotic efficiency implied by the imposition of the constraints under general conditions that allow for nonsmooth objective functions which arise in simulation-based models. We require that the parameters are identified only along

the constrained subspace instead of the whole parameter space. Bayesian credible intervals are asymptotically valid confidence intervals in a pointwise sense, providing exact asymptotic coverage for general functions of the parameters. Our methods encompass the ℓ_1, ℓ_2, and ℓ_∞ penalty functions by defining a kernel function, which appears in part of the quasi-posterior density's limiting density concerning the constrained subspace of the parameter space.

When the constraints are correctly specified, and the penalty parameter diverges at a suitable rate, the posterior distribution along the constraints converges at a faster than $\sqrt{n}$ rate, which induces a singular posterior distribution along the constrained subspace and efficient posterior locations for general nonlinear functions of the parameters that do not lie in the subspace of the constraints. However, if some of the constraints are misspecified, the asymptotic bias can diverge. We therefore consider adaptive methods to identify and place more weight on the correctly specified constraints.

Adaptive methods motivated by the adaptive LASSO of Zou (2006) use the inverse of an initial $\sqrt{n}$-consistent estimator of the constraints to place more weight on the correctly specified constraints and less weight on the misspecified constraints so that the asymptotic bias remains bounded and the posterior mean and median remain $\sqrt{n}$-consistent. Under proper rate restrictions on the penalty parameter, we show that the posterior mean and median can adaptively and selectively identify the correctly specified constraints. Furthermore, Bayesian posterior intervals are asymptotically valid confidence intervals in a pointwise sense, providing asymptotically exact and more efficient coverage for general and possibly nonlinear functions of the parameters.

The prior results apply when the constraints are known to the researcher and depend only on the parameters. If the constraints instead depend on the data, then posterior quantiles cannot be used to form asymptotically valid confidence intervals unless the constraints are asymptotically negligible in the sense that when evaluated at the true parameter value, they converge in probability to zero at a rate faster than $1/\sqrt{n}$. However, a consistent estimate of the constrained estimator's influence function can still be used for asymptotically valid inference.

Related Literature. Alhamzawi et al. (2012), Hans (2009), and Leng et al. (2014) study the Bayesian LASSO and its variants from a fully Bayesian perspective without reference to

frequentist inference. The latter two papers refer to their methods as the "Bayesian adaptive LASSO," which is distinct from the notion of adaptiveness used in Zou (2006) and in this paper.

There is a large literature on the use of unpenalized Laplace-type estimators (LTEs) to simplify the computation of unconstrained extremum estimators with nonlikelihood objective functions that can be nonconcave or discontinuous. For example, Blackwell (1985), Chernozhukov and Hong (2003) and Tian et al. (2007) consider the setting where the dimension of the parameter space is fixed, while Belloni and Chernozhukov (2009) allows for parameters of increasing dimension. None of these papers consider penalized LTEs, which are the focus of this paper. Our paper is also related to the literature on Bayesian estimation of moment condition models, which includes, for example, Schennach (2005), Kitamura and Otsu (2011), Chib et al. (2018), Florens and Simoni (2019), and Gallant (2020b).

Our results only pertain to the pointwise asymptotic properties of penalized LTEs. In a series of papers, Leeb and Pötscher argue forcefully for studying the uniform asymptotics of post-model-selection inference procedures, including the LASSO (e.g. Leeb and Pötscher, 2005, 2008a,b). A general lesson from their work is that consistent model-selection procedures (what they also refer to as "sparse estimators") have arbitrarily large risk, and pointwise asymptotics provide poor approximations to the finite-sample behavior of estimators. We only note that the location functionals considered in this paper, the posterior mean and quantiles, are not consistent model-selection estimators.[1]

Outline. Section 2 contains the main theoretical results for penalized Laplace-type estimators using nonadaptive penalties. We extend the results to adaptive penalties where the constraints are possibly misspecified in section 3. Extensions to estimated and simulated constraints are discussed in section 4.1. Section 4.2 discusses an application of our methods to a GMM setting where we would like to enforce that a subset of the sample moment conditions are zero. Section 4.4 shows that the asymptotic distribution of an ℓ_2 penalized estimator obtained after one Newton-Raphson or Gauss-Newton iteration is equal to the asymptotic distribution of the solution to the original constrained optimization problem.

[1] We thank a referee for pointing out that no method can achieve uniform model selection consistency over $\sqrt{n}$ neighborhoods of zero.

Next, section 5 presentes simulation results on the empirical coverage of posterior quantile based confidence intervals in a constrained IV quantile regression example. The empirical application is presented in section 6. Finally, section 7 concludes. All proofs are stated in Appendix B.

2 Penalized and Laplace-Type Estimators

Extremum estimators $\hat{\theta}$ are typically defined as maximizers of random criterion functions $\hat{Q}_n(\theta)$ in the sense that

$$\hat{Q}_n(\hat{\theta}) \geqslant \sup_{\theta \in \Theta \subset \mathbb{R}^K} \hat{Q}_n(\theta) - o_P\left(n^{-1}\right).$$

It is often useful to incorporate nonlinear constraints for identification or efficiency purposes. An extremum estimator subject to (potentially nonlinear) equality constraints $\bar{\theta}$ instead satisfies

$$\hat{Q}_n\left(\bar{\theta}\right) \geqslant \sup_{\theta \in \bar{\Theta}} \hat{Q}_n(\theta) - o_P\left(n^{-1}\right), \quad \text{where} \quad \bar{\Theta} = \{\theta \in \Theta \colon g(\theta) = 0\}, \tag{1}$$

K is the dimension of θ, $g(\theta) = (g_j(\theta), j = 1, \ldots, J)$ for $J \leqslant K$ encodes constraints on θ, $\bar{\Theta}$ denotes the constrained parameter space, and $\bar{\theta}$ is assumed to exactly satisfy the constraints ($\bar{\theta} \in \bar{\Theta}$).

The asymptotic properties of $\bar{\theta}$ for smoothly differentiable $Q_n(\theta)$ using Lagrange multipliers are extensively developed in Gallant (1987). This paper studies potentially nonsmooth objectives and penalized and Laplace-type versions of $\bar{\theta}$, which have computational advantages. Before introducing these estimators, we discuss some examples of constraints.

Example 1. The maximum rank correlation estimator (Han, 1987; Sherman, 1993) corresponds to

$$\hat{Q}_n(\theta) = \binom{n}{2}^{-1} \sum_{i<j} \left\{1(y_i > y_j)\, 1\left(x_i'\theta > x_j'\theta\right) + 1(y_i < y_j)\, 1\left(x_i'\theta < x_j'\theta\right)\right\}.$$

Because $\hat{Q}_n(\theta) = \hat{Q}_n(\gamma\theta)$ for any $\gamma > 0$, a scale normalization is required to guarantee a unique solution. The typical normalization $\theta_1 = 1$ is not innocuous because it implies that the first regressor is non-zero. The preferred normalization adopted in Han (1987) is $\|\theta\| = 1$, which can be imposed by maximizing $\hat{Q}_n(\theta)$ subject to the constraint $g(\theta) = \|\theta\| - 1 = 0$. However, solving this constrained program in practice can be difficult, which motivates the use of MCMC.

Example 2. Hausman and Woutersen (2014) study a semiparametric duration model, which corresponds to $\theta = (\beta', \delta')'$,

$$\hat{Q}_n(\theta) = -\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{l=1}^{K} \sum_{k=1}^{K} [1(T_i \geqslant l) - 1(T_j \geqslant k)] \, 1(Z_i(l;\beta,\delta) < Z_j(k;\beta,\delta)),$$

where $Z_i(l;\beta,\delta) = \sum_{s=1}^{l} \exp\{X_{is}\beta + \delta_s\}$. In addition to the normalization that $\|\beta\| = 1$, they also impose the constraint $\delta_1 = 0$ in order to normalize the integrated baseline hazard in the first time period, so $g(\theta) = (\|\beta\| - 1, \delta_1)$.

Example 3. Sparsity constraints are commonly imposed, for example in LASSO regression (Tibshirani, 1996), penalized quantile regression (Belloni and Chernozhukov, 2011), and machine-learning methods such as support vector machines (Zhu et al., 2004). In the latter case, for $\kappa > 0$,

$$\hat{Q}_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \max\{\rho_\tau(y_i - x_i'\theta) - \kappa, 0\},$$

where $\rho_\tau(u) = (\tau - 1(u \leqslant 0))u$. Imposing a sparsity constraint on the first $J \leqslant K$ components of θ corresponds to $g_j(\theta) = \theta_j$ for $j = 1, \ldots, J$. Unlike the previous two examples, such constraints are often imposed by the econometrician, despite possible misspecification in the sense that the corresponding components of the true parameter may actually be nonzero. Methods discussed in section 3 will allow for misspecified constraints.

Penalized estimator. An alternative to $\bar{\theta}$ is a *penalized M-estimator* θ^+, which satisfies

$$\bar{Q}_n(\theta^+) \geqslant \sup_{\theta \in \Theta} \bar{Q}_n(\theta) - o_P(n^{-1}) \quad \text{where} \quad \bar{Q}_n(\theta) = \hat{Q}_n(\theta) - \text{penalty}_n(g(\theta)). \tag{2}$$

We consider a general class of penalty functions defined using a J-dimensional kernel function $\kappa(\cdot)$ and penalty parameter λ_n:

$$\text{penalty}_n(g(\theta)) = -\frac{1}{n}\log\kappa\left(\lambda_n\sqrt{n}g(\theta)\right).$$

This paper focuses on ℓ_p penalties for $1 \leqslant p \leqslant \infty$ that are of the form

$$-\log\kappa_p(u) = \sum_{j=1}^{J}|u_j|^p, \quad \text{so that} \quad \text{penalty}_n(g(\theta)) = \frac{\lambda_n^p\sqrt{n}^p}{n}\sum_{j=1}^{J}|g_j(\theta)|^p.$$

The cases of $p = 1, 2, \infty$ correspond respectively to the Laplace, normal, and uniform kernel functions, which we respectively denote

$$\kappa_1(u) = e^{-\sum_{j=1}^{J}|u_j|}, \quad \kappa_2(u) = e^{-\sum_{j=1}^{J}u_j^2}, \quad \text{and} \quad \kappa_\infty(u) = 1(\|u\| \leqslant 1).$$

The ℓ_∞ penalty corresponds to an optimization program in which the equality constraints are relaxed to inequality constraints $\|g(\theta)\| \leqslant \epsilon_n \equiv (\lambda_n\sqrt{n})^{-1}$ and

$$\text{penalty}_n(g(\theta)) = \infty \cdot 1\left(\lambda_n\sqrt{n}\|g(\theta)\| > 1\right).$$

Laplace-type estimators. We focus on Laplace-type estimators consisting of taking the mean and quantiles of a quasi-posterior, which can be computed in practice using MCMC. Given a user-defined prior density function $\pi_0(\theta)$ and penalty parameter λ_n, define the *quasi-prior density function*

$$\pi_n(\theta) \propto \pi_0(\theta)\kappa\left(\lambda_n\sqrt{n}g(\theta)\right) \tag{3}$$

and *quasi-posterior density function*

$$p_\theta(\theta|\mathcal{X}_n) = \frac{\pi_n(\theta)e^{n\hat{Q}_n(\theta)}1(\theta\in\Theta)}{\int_\Theta \pi_n(\theta)e^{n\hat{Q}_n(\theta)}d\theta} = \frac{\pi_0(\theta)e^{n\bar{Q}_n(\theta)}1(\theta\in\Theta)}{\int_\Theta \pi_0(\theta)e^{n\bar{Q}_n(\theta)}d\theta}. \tag{4}$$

Note that (4) is similar to the definition of the quasi-posterior in Chernozhukov and Hong

(2003) except for the addition of $\kappa(\lambda_n\sqrt{n}g(\theta))$, which serves to impose the constraints. For intuition, consider the case where $\kappa(u) = \exp\{-\sum_{j=1}^{J}|u_j|\}$, $\pi_0(\theta)$ is the uniform prior, and $g(\theta)$ is a sparsity constraint (Example 3). Then $\pi_n(\theta)$ is a Laplace prior. If we specify $\hat{Q}_n(\theta)$ as the least-squares objective, then the mode of the quasi-posterior (4) corresponds to the LASSO, while the mean and median correspond to the Bayesian LASSO (Park and Casella, 2008).

Remark 1. In principle, nonbinding inequality constraints can be incorporated into the MCMC routine by multiplying the criterion function with an indicator function for the validity of the inequality constraints, and under pointwise asymptotics, nonbinding constraints do not affect the asymptotic distribution. Also, linear equality constraints, assuming they are correctly specified, can be easily imposed by reparameterizing and reducing the dimension of the parameter space. We therefore focus on potentially nonlinear equality constraints. Because they are nonlinear, we do not directly impose them in the construction of the prior, say by replacing $1(\theta \in \Theta)$ in the quasi-posterior with $1(\theta \in \bar{\Theta})$. This would require specialized, computationally intensive, MCMC algorithms because the constrained parameter space $\bar{\Theta}$ is singular with respect to Lebesgue measure. Such algorithms require a starting value close to the estimator $\bar{\theta}$, and the penalized methods proposed here can provide this starting value (see e.g. Gallant, 2020b).

We study inference on a known scalar function of the true parameter $\phi_0 = \phi(\theta_0)$ of the parameter θ, where $\phi(\cdot)$ is twice continuously differentiable. Denote by ϕ_τ^* the posterior τth quantile of ϕ, which satisfies

$$\int 1(\phi(\theta) \leqslant \phi_\tau^*)\, p(\theta|\mathcal{X}_n)\, d\theta = \tau. \tag{5}$$

A point estimate can be based on $\phi_{1/2}^*$, and an equal-tailed confidence interval of level $1-\tau$ is given by $(\phi_{\tau/2}^*, \phi_{1-\tau/2}^*)$. We provide conditions under which this interval has asymptotically correct coverage in the sense that

$$\liminf_{n\to\infty} P\left(\phi_0 \in \left(\phi_{\tau/2}^*, \phi_{1-\tau/2}^*\right)\right) \geqslant 1-\tau.$$

We will see that if ϕ_0 lies in the constrained space, then coverage is conservative, whereas if it does not, coverage is asymptotically exact. A point estimator for ϕ_0 can also be based on the posterior mean:

$$\phi^* = E\left(\phi\left(\theta\right)|\mathcal{X}_n\right) = \int \phi\left(\theta\right) p\left(\theta|\mathcal{X}_n\right) d\theta. \tag{6}$$

Note that Laplace-type estimators are not Bayesian estimators, since the formula for the quasi-posterior density is missing a Jacobian term that reflects the transformation from the moments to the data (see Gallant, 2020a). In addition, to be considered Bayesian, $e^{n\bar{Q}_n(\theta)}$ must also satisfy a finite-sample normality assumption, which may be violated in practice. This assumption can be checked and remedied using the penalization methods proposed in this paper (see Gallant, 2020a).

2.1 Assumptions

We next state conditions used to prove consistency and asymptotic normality of the penalized and Laplace-type estimators in the next two subsections. We will assume here that constraints are *correctly specified*. This is relevant for Examples 1 and 2 but not 3 because usually sparsity constraints are imposed without *a priori* knowledge about which coefficients are truly zero. In section 3, we consider an adaptive Laplace-type estimator that does not require correct specification of constraints, which is useful for all the examples, in particular Example 3.

Assumption 1. *Θ is a compact subset of $\mathbb{R}^K$ containing $\bar{\Theta}$, and the true parameter θ_0 belongs to the interior of the constrained parameter space $\bar{\Theta}$, where the interior is with respect to the topology of $\bar{\Theta}$.*

The second assumption requires constraints to be smooth.

Assumption 2. *The constraints $g\left(\theta\right)$ are three times continuously differentiable in $\theta \in \Theta$. For $G\left(\theta\right) = \frac{\partial g(\theta)}{\partial \theta'} \in \mathbb{R}^{K\times J}$, $G_0 = G\left(\theta_0\right)$ has rank J.*

The third assumption imposes a weak uniform consistency requirement on $\hat{Q}_n(\theta)$. Note that for many M-estimators, we can let $a_n = \sqrt{n}$.

Assumption 3. *There exists a deterministic and three times continuously differentiable function $Q(\theta)$, and a sequence $a_n \to \infty$, $a_n/n = O(1)$, such that $\sup_{\theta\in\Theta} a_n|\hat{Q}_n(\theta) - Q(\theta)| = O_P(1)$.*

The fourth assumption only requires θ_0 to be identified along the constrained subspace $\bar{\Theta}$ instead of the entire parameter space Θ, which is important for Examples 1 and 2.

Assumption 4. *$Q(\theta) \leqslant Q(\theta_0)$ for all $\theta \in \Theta$, and for all $\delta > 0$, there exists $\epsilon > 0$ such that*

$$\sup_{\|\theta-\theta_0\|\geqslant\delta,\theta\in\bar{\Theta}} Q(\theta) - Q(\theta_0) \leqslant -\epsilon.$$

Finally, to establish the asymptotic distribution of the quasi-posterior distribution we assume the existence of a local quadratic approximation to the possibly nonsmooth objective function.

Assumption 5 (*Local Asymptotic Normality*). *There exists a positive semi-definite matrix H_0 such that*

$$Q(\theta) = Q(\theta_0) - \frac{1}{2}(\theta-\theta_0)' H_0 (\theta-\theta_0) + o\left(\|\theta-\theta_0\|^2\right). \tag{7}$$

In addition, there exists a random sequence Δ_{n,θ_0} such that for

$$R_n(\delta) = \sup_{\|h\|\leqslant\sqrt{n}\delta} \frac{n\hat{Q}_n\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\theta_0) - \Delta'_{n,\theta_0}h - n\left(Q(\theta_0 + h/\sqrt{n}) - Q(\theta_0)\right)}{1+\|h\|^2},$$

(a) $R_n(\delta_n) = o_P(1)$ for any $\delta_n \to 0$, and (b) $\Delta_{n,\theta_0} \xrightarrow{d} N(0,\Omega)$ for some positive semi-definite matrix Ω.

Equation (7) implies that $\partial Q(\theta_0)/\partial\theta = 0$, assuming $Q(\theta)$ is differentiable at θ_0. Assumption 5 implies Assumption 4 in Chernozhukov and Hong (2003) and is satisfied for a broad

class of estimators. When $\hat{Q}_n(\theta)$ is continuously differentiable, Δ_{n,θ_0} is the scaled Jacobian $\sqrt{n}\partial\hat{Q}_n(\theta_0)/\partial\theta$, Ω is the asymptotic variance-covariance matrix of the Jacobian, and H_0 is the negative of the population Hessian $-\partial^2 Q(\theta_0)/\partial\theta\partial\theta'$. If $\hat{Q}_n(\theta)$ is non-smooth, then Δ_{n,θ_0} is the subgradient of $\hat{Q}_n(\theta)$ evaluated at θ_0 and scaled by $\sqrt{n}$, Ω is the asymptotic variance-covariance matrix of Δ_{n,θ_0}, and H_0 is the second derivative of the expected value of the subgradient evaluated at θ_0. For example, for quantile regression, $\hat{Q}_n(\theta) = -n^{-1}\sum_{i=1}^n(\tau - 1(y_i \leqslant x_i'\theta))(y_i - x_i'\theta)$, $\Delta_{n,\theta_0} = n^{-1/2}\sum_{i=1}^n(\tau - 1(y_i \leqslant x_i'\theta_0))x_i$, $\Omega = \tau(1-\tau)E[x_i x_i']$, and $H_0 = E[f_{u|x}(0)x_i x_i']$.

Assumption 6. *(a)* $\exists\beta > 0$ *such that* $a_n\lambda_n^p\sqrt{n}^p/n^{1+\beta} \to \infty$. *(b)* $\lambda_n \to \infty$. *(c)* $\lambda_n/\sqrt{n} \to 0$.

Part (a) is needed for consistency of the penalized estimator θ^+. It ensures that the sample penalized objective is within $o_p(1)$ of the population penalized objective after scaling the objective to ensure the penalty will contribute asymptotically. Part (b) is needed to ensure the penalized estimator is sufficiently close to the constrained estimator so that they share the same asymptotic distribution. Part (c) is additionally needed for the posterior distribution to be sufficiently informative for inference. When $p = \infty$, Assumption 6 can be further weakened to $a_n\lambda_n^p\sqrt{n}^p/n \to \infty$. Example 6 and the discussion after the proof of Theorem 4 illustrate why (c) is required.

The final assumption concerns the function for which we would like to conduct inference.

Assumption 7. *The function* $\phi : \mathbb{R}^K \mapsto \mathbb{R}$ *is twice continuously differentiable.*

2.2 Large-Sample Theory

We first investigate consistency and asymptotic normality of the constrained and penalized estimators $\bar{\theta}$ and θ^+ defined in (1) and (2). Then we prove a Bernstein von-Mises type result and show asymptotic normality of the quasi-posterior distribution and validity of quasi-posterior quantiles for inference.

Heuristically, as long as the penalty parameter λ_n is sufficiently large, the penalized estimator θ^+ should be sufficiently close to the equality constrained M-estimator $\bar{\theta}$ so that

they are both consistent and share the same asymptotic distribution. There is a tradeoff between the rate requirement for λ_n and the sample objective function $\hat{Q}_n(\theta)$. It can be shown that if $\hat{Q}_n(\theta)$ globally identifies θ_0 in Θ, meaning that Assumption 4 holds on the full parameter space Θ rather than the constrained subspace $\bar{\Theta}$, then θ^+ is consistent for any sequence λ_n. On the other hand, if $\hat{Q}_n(\theta)$ only uniquely identifies θ_0 on $\bar{\Theta}$, then consistency of θ^+ requires λ_n to be sufficiently large. We only provide a formal result for the latter case.

Our first result establishes consistency of the constrained and penalized estimators, as well as the posterior mean and quantiles.

Theorem 1. *(a) Under Assumptions 1–4, $\bar{\theta} = \theta_0 + o_P(1)$. (b) If additionally Assumption 6(a) holds, then $\theta^+ = \theta_0 + o_P(1)$. (c) If additionally Assumption 7 holds, then $\phi^*_\tau = \phi(\theta_0) + o_P(1)$ for $\tau \in (0,1)$, and $\phi^* = \phi(\theta_0) + o_P(1)$.*

The next theorem derives the asymptotic distribution of $\bar{\theta}$, allowing H_0 to be possibly singular.

Theorem 2. *Let R be a $K \times (K-J)$ matrix of rank $K-J$ such that $R'G_0 = 0$ and $B \equiv (G_0, R)'$ is nonsingular.[2] Under Assumptions 1–5, if $R'H_0R$ is nonsingular,*

$$\sqrt{n}\left(\bar{\theta} - \theta_0\right) = R\left(R'H_0R\right)^{-1}R'\Delta_{n,\theta_0} + o_P(1). \tag{8}$$

To prove the result in the simple case where H_0 is nonsingular, we first linearize the constraints $g(\theta_0 + \bar{h}/\sqrt{n}) = 0$ for $\bar{h} = \sqrt{n}(\bar{\theta} - \theta_0)$ to obtain $G_0'\bar{h} = o_P(1)$ and apply Assumption 5 to expand $n(\hat{Q}_n(\theta_0 + \bar{h}/\sqrt{n}) - \hat{Q}_n(\theta_0))$. Then to obtain the influence function representation, we use arguments in the proof of Theorem 9.1 of Newey and McFadden (1994) for asymptotically linearizing a constrained GMM estimator. For the case where H_0 is singular, we use a transformation of the parameter space following p. 21 of Amemiya (1985).

[2]Such a matrix always exists and is not necessarily unique (Amemiya, 1985, section 1.4.2).

Remark 2. In smooth models with nonsingular H_0, it is known that

$$\sqrt{n}\left(\bar{\theta}-\theta_0\right)=H_0^{-1}\left(I-G_0\left(G_0'H_0^{-1}G_0\right)^{-1}G_0'H_0^{-1}\right)\Delta_{n,\theta_0}+o_P\left(1\right) \tag{9}$$

(Gallant, 1987). In contrast, (8) only requires the weaker condition that $R'H_0R$ is nonsingular. If H_0 is in fact nonsingular, then the influence functions in (8) and (9) coincide. To see this, let $\gamma_1=R\left(R'H_0R\right)^{-1}R'\Delta_{n,\theta_0}$ and $\gamma_2=H_0^{-1}\left(I-G_0\left(G_0'H_0^{-1}G_0\right)^{-1}G_0'H_0^{-1}\right)\Delta_{n,\theta_0}$. As in Amemiya (1985), we calculate that

$$\begin{pmatrix} R'H_0 \\ G_0' \end{pmatrix}(\gamma_1-\gamma_2)=0.$$

Therefore $\gamma_1=\gamma_2$ if the first matrix is nonsingular. But nonsingularity holds because

$$\begin{pmatrix} R'H_0 \\ G_0' \end{pmatrix}\begin{pmatrix} R & G_0 \end{pmatrix}=\begin{pmatrix} R'H_0R & R'H_0G_0 \\ 0 & G_0'G_0 \end{pmatrix}.$$

The right matrix is nonsingular since nonsingular H_0 implies nonsingular $R'H_0R$. Since (R,G_0) is nonsingular, so is the first matrix on the left, as desired.

Our last result for the penalized and constrained estimators shows the two are asymptotically equivalent.

Theorem 3. *Suppose the conditions of Theorem 2 and Assumptions 6(a) and (b) hold. Then* $\theta^+-\bar{\theta}=o_P(n^{-1/2})$.

To prove the result, we first expand the penalty function locally around $\bar{\theta}$ and expand $n(\bar{Q}_n(\bar{\theta}+B^{-1}un^{-1/2})-\bar{Q}_n(\bar{\theta}))$ using Assumption 5. Then for $u^+=B\sqrt{n}(\theta^+-\bar{\theta})$ and $B=(G_0,R)'$, we use the fact that $n(\bar{Q}_n(\bar{\theta}+B^{-1}u^+n^{-1/2})-\bar{Q}_n(\bar{\theta}))\geqslant o_p(1)$ by definition of θ^+ to argue that $u^+=o_p(1)$.

Laplace-type estimators. To derive the asymptotic distribution of Laplace-type estimators, a key step is to prove a generalized Bernstein-von Mises (BvM) result on convergence

of the quasi-posterior density (4). In the typical BvM setting, the influence of the prior disappears, so the density is asymptotically normal. Our setting is different due to the scaling λ_n in $\pi_n(\theta)$, which ensures the prior plays a nontrivial role in the limit. This is needed to impose the constraints.

A common technique for deriving a BvM result is to study the quasi-posterior of a localized parameter (e.g. van der Vaart, 2000). Let $h = \sqrt{n}\left(\theta - \bar{\theta}\right)$, and define the localized quasi-posterior density

$$p_h\left(h \mid \mathcal{X}_n\right) = \frac{p_\theta(\bar{\theta} + h/\sqrt{n} \mid \mathcal{X}_n)}{\sqrt{n}^K} = \frac{\pi_n(\bar{\theta} + h/\sqrt{n}) e^{n\hat{Q}_n(\bar{\theta}+h/\sqrt{n})} 1(h \in \sqrt{n}(\Theta - \bar{\theta}))}{\int_{h \in \sqrt{n}(\Theta-\bar{\theta})} \pi_n(\bar{\theta} + h/\sqrt{n}) e^{n\hat{Q}_n(\bar{\theta}+h/\sqrt{n})} dh}. \tag{10}$$

Note that h is defined by localizing relative to the constrained estimator $\bar{\theta}$, rather than the true parameter θ_0, which is more typically done for unconstrained models (e.g. Chernozhukov and Hong, 2003).

Parameters inside the constrained subspace will have different asymptotic behavior than those outside. To handle this, we first transform the localized parameter space in a manner similar to Theorem 2. As in that theorem, let R be a $K \times (K - J)$ matrix of rank $K - J$ such that $R'G_0 = 0$. For $B = (G_0, R)'$ and $u = Bh$, let

$$p_u\left(u \mid \mathcal{X}_n\right) = \frac{1}{|\det\left(B\right)|} p_h\left(B^{-1}u \mid \mathcal{X}_n\right).$$

Now partition u into two parts: u_1, which contains the first J components of u and corresponds to the parameters contained inside the constrained subspace, and u_2, which contains the remaining $K - J$ components of U and corresponds to parameters off the constrained subspace. Let $v = (v_1', v_2')' = (\lambda_n u_1', u_2')' = D_n u$ with $D_n = \operatorname{diag}\left(\lambda_n I_J, I_{K-J}\right)$. Define the reparametrized, localized quasi-posterior

$$p_v\left(v \mid \mathcal{X}_n\right) = \lambda_n^{-|J|} p_u\left(D_n^{-1}v \mid \mathcal{X}_n\right) = \frac{1}{\lambda_n^{|J|}|\det\left(B\right)|} p_h\left(B^{-1}D_n^{-1}v \mid \mathcal{X}_n\right). \tag{11}$$

This corresponds to a $\lambda_n\sqrt{n}$ scaling for parameters contained inside the constrained subspace and a $\sqrt{n}$ scaling for parameters off the constrained subspace, which reflects their differing rates of convergence.

We can now state our BvM result. Define the total variation of moments norm $\|f(\cdot) - g(\cdot)\|_\alpha = \int \|h\|^\alpha |f(h) - g(h)| dh$ for $\alpha \geqslant 0$ and densities f, g. The usual BvM theorem shows convergence under the total variation norm, which corresponds to our norm with $\alpha = 0$. As in Chernozhukov and Hong (2003), we will need the stronger norm in order to guarantee convergence of posterior moments in a later result.

Theorem 4. *Let Assumptions 1–6 hold. Then for any* $0 \leqslant \alpha < \infty$,[3]

$$\|p_v(\cdot|\mathcal{X}_n) - p_\infty(\cdot)\|_\alpha = o_P(1)$$

for $p_\infty(v) = p_{1\infty}(v_1) p_{2\infty}(v_2)$, *where*

$$p_{1\infty}(v_1) = \begin{cases} C_\kappa^{-1} e^{-\sum_{j=1}^J |v_{1j}|^p} & \text{for } p < \infty \\ C_\kappa^{-1} 1(\|v_1\| \leqslant 1) & \text{for } p = \infty \end{cases}, \quad p_{2\infty}(v_2) = \frac{\det(\Sigma)^{-1/2}}{\sqrt{2\pi}^{K-J}} e^{-\frac{1}{2} v_2' \Sigma^{-1} v_2},$$

$\Sigma^{-1} = (R'R)^{-1} R' H_0 R (R'R)^{-1}$, *and*

$$C_\kappa = \begin{cases} \left(\int e^{-|u|^p} du\right)^J & \text{for } p < \infty \\ \frac{1}{2^J} & \text{for } p = \infty \end{cases}$$

This shows the quasi-posterior density of v converges in the total variation of moments norm to a product of two densities. One part is a multivariate mean zero normal random vector corresponding to the unconstrained part v_2 and the other a density given by the kernel function corresponding to the constrained part v_1. The quasi-posterior density concentrates around the constrained estimator $\bar{\theta}$ at a $\sqrt{n}$ rate off the constraints but at a $\lambda_n \sqrt{n}$ rate along the constraints.

To prove the result, we apply Assumption 5 to $n(\hat{Q}_n(\bar{\theta} + B^{-1} D_n^{-1} v/\sqrt{n}) - \hat{Q}_n(\bar{\theta}))$ and

[3]Our result applies for α fixed with respect to n. See e.g. Theorem 2.2 of Belloni and Chernozhukov (2014) for a result in a setting without constraints where α can diverge.

linearize the constraints $g(\bar{\theta} + B^{-1}D_n^{-1}v/\sqrt{n})$. We then reduce the problem to showing (68):

$$\int_{\|v\| \geqslant M_n, v \in H_n, \|B^{-1}D_n^{-1}v\| \leqslant \sqrt{n}\delta} \|v\|^{\alpha} \hat{\pi}_0(v) \exp(w(v)) \, dv = o_P(1)$$

for any $\delta \to 0$, where $\hat{\pi}_0(v) = \pi_0(\bar{\theta} + B^{-1}D_n^{-1}v/\sqrt{n})$, $H_n = D_n B\sqrt{n}(\Theta - \bar{\theta})$, and $w(v) = n(\hat{Q}_n(\bar{\theta} + B^{-1}D_n^{-1}v/\sqrt{n} - \hat{Q}_n(\bar{\theta})) + \sum_{j=1}^{J} |\lambda_n \sqrt{n} g_j(\bar{\theta} + B^{-1}D_n^{-1}v/\sqrt{n})|^p$.

Remark 3. It is instructive to compare the limit distribution $p_\infty(\cdot)$ in Theorem 4 with the limit in the unconstrained case. The latter is given in Theorem 1 of Chernozhukov and Hong (2003), namely the multivariate normal density with mean zero and variance H_0. Denote this density by $\tilde{p}_\infty(h)$. That theorem implies that $h \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} \tilde{p}_\infty(\cdot)$, meaning

$$\sup_x \left| \int^x p_h(h \mid \mathcal{X}_n) \, dh - \int^x \tilde{p}_\infty(h) \, dh \right| = o_P(1).$$

Now, Theorem 4 implies that $v \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} p_\infty(\cdot)$. Recalling that $v = (\lambda_n u_1' \;\; u_2')'$, this has two consequences of note. First, $u_1 = v_1/\lambda_n = o_P^*(1)$, by which we mean that $\forall \epsilon > 0$, $\int_{\|u_1\| \geqslant \epsilon} p_u(u \mid \mathcal{X}_n) \, du = o_P(1)$. This follows from the fact that

$$\begin{aligned}\int_{\|u_1\| \geqslant \epsilon} p_u(u \mid \mathcal{X}_n) \, du &= \int_{\|v_1\| \geqslant \lambda_n \epsilon} p_v(v \mid \mathcal{X}_n) \, du = \int_{\|v_1\| \geqslant \lambda_n \epsilon} p_\infty(v_1) \, dv_1 + o_P(1) \\ &\leqslant \sum_{j \in J} \int_{\|v_{1j}\| \geqslant \frac{\lambda_n \epsilon}{J}} p_\infty(v_{1j}) \, dv_{1j} + o_P(1) = \sum_{j \in J} e^{-\frac{\lambda_n \epsilon}{J}} + o_P(1) = o_P(1).\end{aligned}$$

Second, $u_2 \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} N(0, \Sigma)$. Finally, recalling that $h = B^{-1}(u_1', u_2')'$, by the Bootstrap CMT (Kosorok, 2007, Proposition 10.7),

$$\begin{aligned}h \equiv B^{-1}u =& R(R'R)^{-1} u_2 + G_0 (G_0'G_0)^{-1} u_1 \\ =& R(R'R)^{-1} u_2 + o_P^*(1) \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} R(R'R)^{-1} N(0, \Sigma) = RN\left(0, (R'H_0R)^{-1}\right).\end{aligned} \tag{12}$$

Therefore, $p_h(\cdot \mid \mathcal{X}_n)$ converges to the density of a singular multivariate normal distribution described on p. 32 of Anderson (1958).

Inference on scalar functions of parameters can be based on (5) and (6) using the Delta method. The next theorem shows that the asymptotic normality of the posterior distribution established in Theorem 4 translates into desirable statistical properties of the MCMC computational procedure. First, the posterior mean and median of scalar functions $\phi(\cdot)$ of θ are $\sqrt{n}$-consistent and asymptotically equivalent to $\phi\left(\bar{\theta}\right)$. Second, confidence intervals constructed from quasi-posterior quantiles have correct asymptotic coverage.

Theorem 5. *Let Assumptions 1 to 7 hold. Let* $\Lambda = \frac{\partial \phi(\theta_0)}{\partial \theta}$. *Then* ϕ^* *and* $\phi^*_{1/2}$ *are asymptotically equivalent to* $\phi\left(\bar{\theta}\right)$*:*

$$\phi^* - \phi\left(\bar{\theta}\right) = o_P(n^{-1/2}) \quad \textit{and} \quad \phi^*_{1/2} - \phi\left(\bar{\theta}\right) = o_P(n^{-1/2}). \tag{13}$$

Also, for any $\tau \in (0,1)$*,*

$$\phi^*_\tau - \phi\left(\bar{\theta}\right) - q_\tau \frac{1}{\sqrt{n}} \sqrt{\Lambda' R \left(R' H_0 R\right)^{-1} R' \Lambda} = o_P(n^{-1/2}). \tag{14}$$

where $q_\tau = \Phi^{-1}(\tau)$ *is the* τ*th quantile of the standard normal distribution.*

Furthermore, if $\Lambda' R \neq 0$ *and the "information matrix equality"* $R'\Omega R = R' H_0 R$ *holds, where* Ω *and* H_0 *are defined in Assumption 5, then posterior quantile confidence intervals are asymptotically exact:*

$$\lim_{n\to\infty} P\left(\phi_0 \in \left(\phi^*_{\tau/2}, \phi^*_{1-\tau/2}\right)\right) = 1 - \tau \tag{15}$$

On the other hand, if $\Lambda = G_0' \eta$ *for some* η *(so that* $\Lambda' R = 0$*), then*

$$\lambda_n \sqrt{n}\left(\phi^* - \phi\left(\bar{\theta}\right)\right) = o_P(1) \quad \textit{and} \quad \lambda_n \sqrt{n}\left(\phi^*_\tau - \phi\left(\bar{\theta}\right)\right) = \bar{q}_\tau + o_P(1), \tag{16}$$

where $\bar{q}_\tau$ *is the* τ*-th quantile of* $\eta' V_1$ *and* V_1 *is distributed as* $p_{1\infty}(v_1)$*, and posterior quantile confidence intervals are asymptotically conservative:*

$$\lim_{n\to\infty} P\left(\phi_0 \in \left(\phi^*_{\tau/2}, \phi^*_{1-\tau/2}\right)\right) \geqslant 1 - \tau. \tag{17}$$

If $\Lambda = G_0'\eta$ for some η, this is the setting in which the parameters lie along the constraints. The theorem shows that the quasi-posterior mean and median are then superconsistent, and confidence intervals obtained from quasi-posterior quantiles are conservative. If $\Lambda' R \neq 0$, the parameters do not lie along the constraints, so the Laplace-type estimators are $\sqrt{n}$-consistent, and the quasi-posterior quantiles can be used to obtain asymptotically exact coverage.

The key arguments in the proof establish the following conditional Delta method result: $\sup_{s\in R}|F_{n,\phi}(s) - F_{\phi,\infty}(s)| = o_P(1)$. In the case where parameters do not lie along the constraints, $F_{n,\phi}(s) = P(\sqrt{n}(\phi(\theta) - \phi(\bar{\theta})) \leqslant s \,|\, \mathcal{X}_n)$ is the quasi-posterior distribution of $\phi(\theta)$ (centered and scaled), with limit $F_{\phi,\infty}(s) = \int_{\Lambda' R(R'R)^{-1} v_2 \leqslant s} p_v^{\infty}(v)\, dv = \Phi(s/\sqrt{\Lambda' R (R' H_0 R)^{-1} R'\Lambda})$. Thus, we obtain a normal limit, and as in Chernozhukov and Hong (2003), valid posterior quantile confidence intervals under the information matrix equality. On the other hand, in the case where parameters lie along the constraints, $F_{n,\phi}(s) = P(\lambda_n\sqrt{n}(\phi(\theta) - \phi(\bar{\theta})) \leqslant s \,|\, \mathcal{X}_n)$, which has a different scaling of $\lambda_n\sqrt{n}$. For the case of the ℓ_1 penalty, the corresponding limit is $F_{\phi,\infty}(s) = \int_{\eta' v_1} p_v^{\infty}(v)\, dv = \int_{\eta' v_1} \left(\frac{1}{2}\right)^J \prod_{j\in J} e^{-|v_{1j}|} dv_1$, which is the Laplace density. Notably, the latter does not depend on H_0, so the result holds regardless of whether we obtain the information matrix equality.

Remark 4 (Vector functions)**.** When $\phi(\theta)$ is a vector, $\Lambda' R \neq 0$, and Λ is not linearly dependent with $R'\Omega R = R' H_0 R$, (13) continues to hold. Similar to $\phi(\bar{\theta})$, $\sqrt{n}(\phi^* - \phi(\theta_0))$ and $\sqrt{n}(\phi^*_{1/2} - \phi(\theta_0))$ are both asymptotically $N(0, \Lambda' R (R' H_0 R)^{-1} R'\Lambda)$. The quasi-posterior joint distribution of $\phi(\theta)$ can be used to estimate the asymptotic variance matrix consistently. Apply (14) to any linear combination $\eta_l'\phi(\theta), l = 1, \ldots, L$,

$$\eta_l'\Lambda' R (R' H_0 R)^{-1} R'\Lambda\eta_l = n\left(\frac{\eta_l'\left(\phi^*_{\tau_{l1}} - \phi^*_{\tau_{l2}}\right)}{q_{\tau_{l1}} - q_{\tau_{l2}}}\right)^2 + o_P(1).$$

All elements of $\Lambda' R (R' H_0 R)^{-1} R'\Lambda$ can then be estimated consistently by varying η_l and τ_{l1} and τ_{l2}. Alternatively, the joint posterior variance-covariance matrix of $\phi(\theta)$ also estimates

the asymptotic variance consistently:

$$\begin{aligned} n \operatorname{Var}(\phi(\theta) \mid \mathcal{X}_n) &\equiv n \int (\phi(\theta) - \phi^*)(\phi(\theta) - \phi^*)' p(\theta \mid \mathcal{X}_n) \, d\theta \\ &= \Lambda' R (R' H_0 R)^{-1} R' \Lambda + o_P(1). \end{aligned} \tag{18}$$

This equation is established at the end of the proof of Theorem 5.

3 Adaptation to Misspecified Constraints

Thus far we have assumed that all constraints are correctly specified (Assumption 4). However, if some are misspecified, then the asymptotic bias of the estimators in the previous section can diverge, which motivates the use of adaptive methods. In this section, we adopt the strategy of the adaptive LASSO (Zou, 2006), which is to use a preliminary estimator to reweight the constraints such that in the limit, the penalty term vanishes for the misspecified constraints, leaving only the penalty on the correctly specified constraints.

Let $(g_m(\theta), m = 1, \ldots, J)$ be the J correctly specified constraints, $(g_m(\theta), m = J+1, \ldots, M)$ the $L = M - J$ misspecified constraints, and $g(\theta)$ the vector of all M constraints. We do not assume knowledge of J. We recycle the notation from section 2, first redefining the constrained parameter space

$$\bar{\Theta} = \{\theta \in \Theta \colon g_m(\theta) = 0 \ \ \forall m = 1, \ldots, J\}. \tag{19}$$

The true parameter θ_0 is now defined as satisfying Assumption 4 but for $\bar{\Theta}$ in (19).

Let $\tilde{\theta}$ be a preliminary $\sqrt{n}$-consistent estimate of θ_0. Define $\hat{w} = (\hat{w}_m, m = 1, \ldots, M)$, a vector of data-dependent weights, with $\hat{w}_m = |g_m(\tilde{\theta})|^{-\gamma}$ for some $\gamma > 0$ and all m. We define the adaptively penalized objective function as

$$\bar{Q}_n(\theta) = \hat{Q}_n(\theta) - \text{penalty}_n(g(\theta)) \quad \text{for} \quad \text{penalty}_n(g(\theta)) = -\frac{1}{n} \log \kappa_p(\lambda_n \hat{w} \circ g(\theta)) \tag{20}$$

As in the adaptive LASSO, the idea is that, for any constraint m that is misspecified, the associated weight $\hat{w}_m$ converges in probability to a positive constant, and our conditions on λ_n

ensure that the constraint does not asymptotically contribute. In contrast, if the constraint is correctly specified, then $\hat{w}_m = O_p(\sqrt{n}^{\gamma})$, which boosts the penalty on the constraint so that it matters in the limit.

As before, we focus on the ℓ_p penalties of the form of

$$-\log \kappa_p(u) = \sum_{m=1}^{M} |u_m|^p \quad \text{so that} \quad \text{penalty}_n(g(\theta)) = \frac{\lambda_n^p}{n} \sum_{m=1}^{M} |\hat{w}_m g_m(\theta)|^p.$$

The cases of $p = 1, 2$ correspond respectively to the Laplace and normal kernel functions:

$$\kappa_1(u) = e^{-\sum_{m=1}^{M} |u_m|} \implies \text{penalty}_n(g(\theta)) = \frac{\lambda_n}{n} \sum_{m=1}^{M} \hat{w}_m |g_m(\theta)|,$$

$$\kappa_2(u) = e^{-\sum_{m=1}^{M} u_m^2} \implies \text{penalty}_n(g(\theta)) = \frac{\lambda_n^2}{n} \sum_{m=1}^{M} \hat{w}_m^2 g_m(\theta)^2.$$

Unlike section 2, we now need to restrict $p \in [1, \infty)$, in particular ruling out the ℓ_∞ case. To see why, note that this penalty corresponds to

$$\kappa_\infty(u) = 1(\|u\| \leqslant 1) \implies \text{penalty}_n(g(\theta)) = \infty \cdot 1\left(\sum_{m=1}^{M} \hat{w}_m |g_m(\theta)| > \frac{1}{\lambda_n}\right). \tag{21}$$

For the misspecified constraints, our conditions will imply that the associated weight $\hat{w}_m$ converges to a positive constant, while $\lambda_n \to \infty$. Then since the constraint is misspecified ($g_m(\theta_0) \neq 0$), eventually the estimator infinitely penalizes all parameters that fail to satisfy the misspecified constraints, including the true parameter, which is the opposite of what needs to happen. This also suggests that larger values of p can end up penalizing too much based on misspecified constraints, so to compensate, we will need to ensure that the penalty parameter λ_n diverges more slowly when p is larger. A new condition below formalizes this requirement.

Estimators. Let us redefine the constrained estimator $\bar{\theta}$ using the new constrained subspace (19). Note that this is an "oracle" estimator because it uses only the constraints that are correctly specified. Its consistency follows from Theorem 1(a).

In practice, we consider adaptive estimators that do not require knowledge of the correctly

specified constraints. Toward that end, we redefine the penalized estimator θ^+ from section 2 using the above adaptive weights. For the Laplace-type estimators, we redefine the quasi-prior using adaptive weights as follows:

$$\pi_n(\theta) \propto \pi_0(\theta)\, \kappa_p(\lambda_n \hat{w} \circ g(\theta)) \tag{22}$$

Example 4. Consider the case where $\pi_0(\theta)$ is uniform, $\hat{Q}_n(\theta)$ is the least-squares objective, and $g(\theta)$ imposes the sparsity constraints of Example 3. If $p = 1$, then the mode of the quasi-posterior corresponds to the adaptive LASSO of Zou (2006). Taking the mean or median instead, we obtain adaptive versions of the Bayesian LASSO. If $p = 2$, then the mode is an adaptive version of ridge regression.

3.1 Assumptions

To construct the weights $\hat{w}$, we require a preliminary estimator $\tilde{\theta}$ that is $\sqrt{n}$-consistent for θ_0 in the following sense.

Assumption 8. *For all $m = 1, \ldots, M$ and $\tilde{g}_m \equiv g_m(\tilde{\theta})$, $\sqrt{n}\,(\tilde{g}_m - g_m(\theta_0)) = O_P(1)$.*

In practice, if θ_0 is globally identified, then $\tilde{\theta}$ can be obtained from the unconstrained estimator that maximizes $\hat{Q}_n(\theta)$.

Next, we require the penalty parameter to satisfy the following rate conditions.

Assumption 9. *Let $\bar{\lambda}_n = \lambda_n n^{\frac{\gamma-1}{2}}$, where $\gamma > 0$ is used in the definition of the weights $\hat{w}_m = |\tilde{g}_m|^{-\gamma}$. (a) $\exists \beta > 0$ such that $a_n \bar{\lambda}_n^p \sqrt{n^p}/n^{1+\beta} \to \infty$. (b) $\bar{\lambda}_n \to \infty$. (c) $\bar{\lambda}_n/\sqrt{n} = o(1)$. (d) $\lambda_n^p/\sqrt{n} = o(1)$.*

Conditions (a)–(c) are analogous to Assumption 6 for the non-adaptive estimator. Condition (d) is new and reflects the intuition from the discussion following (21). To see that this is still compatible with (a), note that in the typical case where $a_n = \sqrt{n}$, (a) is equivalent to $\frac{\lambda_n^p}{\sqrt{n}} \frac{n^{\gamma p/2}}{n^\beta}$, so the first fraction degenerating to zero is compatible with the product of the fractions diverging, since β can be chosen arbitrarily small.

Strictly speaking, to allow for the possibility that $\tilde{g}_m = 0$ and to construct a nondegenerate posterior density, we should redefine

$$\hat{w}_m = \frac{1}{|\tilde{g}_m|^\gamma} 1(\tilde{g}_m \neq 0) + \sqrt{n}^\gamma 1(\tilde{g}_m = 0).$$

In theory, if $\tilde{g}_m = 0$ for all $m \in \bar{M}$, and either the constraints are linear in the parameters or the nonlinear constraints can be inverted, the quasi-posterior distribution can be redefined to place probability 1 on the restricted parameter set: $g_m(\theta) = 0, m \in \bar{M}$:

$$p_\theta(\theta|\mathcal{X}_n) = \frac{\pi_0(\theta) e^{n\hat{Q}_n(\theta) - \lambda_n \sum_{m \in \bar{M}^c} \hat{\omega}_m |g_m(\theta)|} 1\left(\theta : g_m(\theta) = 0, \forall m \in \bar{M}\right)}{\int_{\theta: g_m(\theta)=0, \forall m \in \bar{M}} \pi_0(\theta) e^{n\hat{Q}_n(\theta) - \lambda_n \sum_{m \in \bar{M}^c} \hat{\omega}_m |g_m(\theta)|} d\theta} \tag{23}$$

However, direct implementation of (23) can be difficult for general nonlinear constraints (see Remark 1).

Assumption 10. *For all $m \in M$, $\left(\sqrt{n}\left(\tilde{g}_m - g_m(\theta_0)\right)\right)^{-1} = O_P(1)$.*

The assumption states that $\tilde{g}_m$ is consistent at exactly the $\sqrt{n}$ rate. This is analogous to the assumption in Zou (2006) that the exact rate of the initial estimate is known. In principle, we could allow for other rates, provided we adjust the rate conditions on λ_n appropriately.

3.2 Large-Sample Theory

The first theorem establishes asymptotic equivalence of the adaptively penalized and constrained estimators.

Theorem 6. *Suppose the conditions of Theorem 2 hold (using the new definition of the constrained space (19)). Under Assumptions 8 and 9, $\theta^+ = \theta_0 + o_P(1)$ and $\theta^+ - \bar{\theta} = o_P(n^{-1/2})$.*

We next provide an analog of Theorems 4 and 5 for adaptive Laplace-type estimators. Redefine the localized quasi-posterior $p_v(v \mid \mathcal{X}_n)$ from (11) by replacing λ_n with $\bar{\lambda}_n$ in As-

sumption 9. That is, let $\bar{D}_n = \text{diag}\left(\bar{\lambda}_n I_J, I_{K-J}\right)$ and

$$p_v\left(v \mid \mathcal{X}_n\right) = \frac{1}{\bar{\lambda}_n^J |\det\left(B\right)|} p_h\left(B^{-1}\bar{D}_n^{-1} v \mid \mathcal{X}_n\right),$$

where $p_h(\cdot \mid \mathcal{X}_n)$ is now defined using the adaptive quasi-prior (22).

Theorem 7. *Under Assumptions 1–10, the conclusions of Theorems 4 and 5 hold with* $p_\infty\left(v\right)$ *replaced by* $p_\infty\left(v \mid \mathcal{X}_n\right) = p_{1\infty}\left(v_1 \mid \mathcal{X}_n\right) p_{2\infty}\left(v_2\right)$*, where*

$$p_{1\infty}\left(v_1 \mid \mathcal{X}_n\right) = \bar{C}_{\kappa_J}^{-1} e^{-\sum_{j=1}^J |\sqrt{n}\tilde{g}_j|^{-p\gamma}|v_{1j}|^p}, \quad p_{2\infty}\left(v_2\right) = \frac{\det\left(\Sigma\right)^{-1/2}}{\sqrt{2\pi}^{K-J}} e^{-\frac{1}{2}v_2'\Sigma^{-1}v_2},$$

$\Sigma^{-1} = \left(R'R\right)^{-1} R'H_0 R \left(R'R\right)^{-1}$*, and* $\bar{C}_{\kappa_J} = \int e^{-\sum_{j=1}^J |\sqrt{n}\tilde{g}_j|^{-p\gamma}|v_{1j}|^p} du$*, and (16) replaced by*

$$\bar{\lambda}_n\sqrt{n}\left(\phi^* - \phi\left(\bar{\theta}\right)\right) = o_P\left(1\right) \quad \textit{and} \quad \bar{\lambda}_n\sqrt{n}\left(\phi_\tau^* - \phi\left(\bar{\theta}\right)\right) = \bar{q}_\tau^* + o_P\left(1\right),$$

where $\bar{q}_\tau^*$ *is the* τ*-th quantile of* $\eta' V_1$ *and* V_1 *is distributed as* $p_{1\infty}\left(v_1 \mid \mathcal{X}_n\right)$*.*

This theorem shows that the quasi-posterior density concentrates around the constrained estimator $\bar{\theta}$ at a $\sqrt{n}$ rate off the constraints but at a $\bar{\lambda}_n\sqrt{n}$ rate along the correctly specified constraints. When $\phi\left(\theta_0\right)$ lies along the correctly specified constraints, ϕ^* and $\phi_{1/2}^*$ are superconsistent, and equal-tailed quasi-posterior credible intervals provide asymptotically conservative coverage. On the other hand, if, say, all constraints are misspecified, then we still obtain $\sqrt{n}$-consistency, and the credible interval provides asymptotically exact coverage. This is because the misspecified constraints do not enter the limiting posterior, as the adaptive weighting removes the impact of the part of the quasi-prior involving $(g_m(\theta), m = J+1, \ldots, M)$. Thus, in the case where all constraints are misspecified, the limiting quasi-posterior is simply normal, which is the usual BvM result.

Remark 5. As discussed in Example 4, the class of Laplace-type estimators we consider contains adaptive versions of the Bayesian LASSO. For the ℓ_1 penalty, it is well-known that the mode of the quasi-posterior leads to concurrent model selection and shrinkage, as in the

LASSO. The previous theorem indicates that the quasi-posterior mean or median, as in the Bayesian LASSO, have no model selection properties. More precisely, the posterior mode can have an asymptotic point mass at zero in certain dimensions of the parameter space, which is what is meant by model selection (Knight and Fu, 2000). In contrast, it can be shown, using our characterization of the limiting quasi-posterior, that the posterior mean and median have no such point mass. This is also true for the non-adaptive case in section 2.

Example 5. We illustrate the contrast between our adaptive and non-adaptive Laplace-type estimators for the case of estimating the mean. To obtain simple expressions for the quasi-posterior, we consider the ℓ_2 penalty. Let $\pi_0(\theta)$ be the uniform prior, and consider the constraint $g(\theta) = \theta^*$. Then $\pi_n(\theta) \sim N(\theta^*, n^{-1}\lambda_n^{-2})$. Suppose that $X_i \sim N(\theta, \sigma^2)$, so that $\bar{X} \equiv n^{-1}\sum_{i=1}^n X_i \sim N(\theta, \sigma^2 n^{-1})$. Then the posterior distribution of θ is

$$\theta \mid \mathcal{X}_n \sim N\left(\frac{\frac{n}{\sigma^2}\bar{X} + n\lambda_n^2\theta^*}{\frac{n}{\sigma^2} + n\lambda_n^2}, \frac{1}{\frac{n}{\sigma^2} + n\lambda_n^2}\right).$$

Recall that θ_0 denotes the true parameter, which may differ from θ^*. Write $X_n \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} Y_n$ if $\rho_{BL_1}(X_n, Y_n) = o_P(1)$, where $\rho_{BL_1}(\cdot)$ metrizes weak convergence.

First consider the non-adaptive prior. Under correct specification, where $\theta^* = \theta_0$, if $\lambda_n \to \infty$, then

$$\lambda_n\sqrt{n}(\theta - \theta_0) \mid \mathcal{X}_n \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} N(0, 1).$$

In contrast, under misspecification, where $\theta^* \neq \theta_0$,

$$\lambda_n\sqrt{n}(\theta - \theta_0) \mid \mathcal{X}_n \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} N\left(\lambda_n\sqrt{n}(\theta^* - \theta_0), 1\right).$$

Therefore, the asymptotic bias diverges under misspecification.

Next consider the adaptive prior using $\bar{X}$ as the initial estimate. Then the estimated constraint is $\tilde{g} = \bar{X} - \theta^*$, and the adaptive prior is $\pi_n(\theta) \sim N(\theta^*, (\bar{X} - \theta^*)^2\lambda_n^{-2})$, resulting

in the posterior

$$\theta \mid \mathcal{X}_n \sim N\left(\frac{\frac{n}{\sigma^2}\bar{X} + \left(\bar{X} - \theta^*\right)^{-2}\lambda_n^2\theta^*}{\frac{n}{\sigma^2} + \left(\bar{X} - \theta^*\right)^{-2}\lambda_n^2}, \frac{1}{\frac{n}{\sigma^2} + \left(\bar{X} - \theta^*\right)^{-2}\lambda_n^2}\right).$$

Now under correct specification, if $\lambda_n \to \infty$, then

$$\lambda_n\sqrt{n}\left(\theta - \theta_0\right) \mid \mathcal{X}_n \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} N\left(0, n\left(\bar{X} - \theta_0\right)^2\right).$$

Note that the variance differs from the non-adaptive case due to the randomness in the estimated constraint $\tilde{g}$. In contrast, under misspecification, if $\lambda_n/\sqrt{n} \to 0$, then

$$\sqrt{n}\left(\theta - \theta_0\right) \mid \mathcal{X}_n \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} N\left(\sqrt{n}\left(\bar{X} - \theta_0\right), \sigma^2\right).$$

This is the limit we would obtain in the standard unconstrained case. In particular, the asymptotic bias remains stochastically bounded, unlike the non-adaptive case, and the rate of convergence is $\sqrt{n}$ instead of $\lambda_n\sqrt{n}$.

4 Generalizations

In this section we discuss several generalizations. They are far from exhaustive and only serve to illustrate the scope of additional directions.

4.1 Estimated and Simulated Constraints

In empirical applications sometimes the constraints $g\left(\theta\right)$ can only be estimated or simulated by some $g_n\left(\theta\right)$. The next assumption allows for either. For example, we can allow for simulated constraints

$$g_n\left(\theta\right) = S\left(n\right)^{-1}\sum_{j=1}^{S(n)} g\left(\xi_j, \theta\right),$$

where the ξ_j's represent simulation draws and $S\left(n\right)$ denotes the number of simulations, which depends on the sample size. Constraints can also be estimated from sample data:

$g_n(\theta) = P_n g(\cdot, \theta)$ where P_n is the empirical measure.

Assumption 11. *(a)* $\sup_{\theta \in \Theta} \|g_n(\theta) - g(\theta)\| = o_P(1)$. *(b)* $\sqrt{n} g_n(\theta) = O_p(1)$. *(c)*

$$\sup_{\|\theta - \theta_0\| \leqslant o(1)} \sqrt{n} \left(g_n(\theta) - g(\theta) - g_n(\theta_0) + g(\theta_0) \right) = o_P(1). \tag{24}$$

While $g(\theta)$ is required to be smooth, $g_n(\theta)$ can be discontinuous, which makes it difficult for $g_n(\theta)$ to be exactly zero in finite sample. We therefore relax the constraint to $\|g_n(\bar{\theta}_S)\| \leqslant \epsilon_n$, for $\epsilon_n = o_P(n^{-1/2})$, and define the constrained estimator using estimated/simulated constraints $\bar{\theta}_S$ as $\hat{Q}_n(\bar{\theta}_S) \geqslant \sup_{\theta \in \Theta : \|g_n(\theta)\| \leqslant \epsilon_n} \hat{Q}_n(\theta) - o_P(n^{-1})$. The next theorem demonstrates that $\bar{\theta}_S$ is consistent and gives its influence function representation.

Theorem 8. *Under Assumptions 1–5, and 11,*

$$\begin{aligned} \sqrt{n}\left(\bar{\theta}_S - \theta_0\right) =& R\left(R' H_0 R\right)^{-1} R' \Delta_{n,\theta_0} \\ & - \left(I - R\left(R' H_0 R\right)^{-1} R' H_0\right) G_0 \left(G_0' G_0\right)^{-1} \sqrt{n} g_n(\theta_0) + o_P(1). \end{aligned} \tag{25}$$

Remark 6. Note that $\bar{\theta}_S$ has an influence function that differs from the one in (8) due to the presence of the additional second term in (25), which captures the additional variation from the estimated constraints. Unless these constraints satisfy $\mathrm{Var}(g_n(\theta)) = 0$, or they are asymptotically negligible in the sense that $\sqrt{n} g_n(\theta_0) = o_P(1)$, then the information matrix equality will generally not hold and quasi-posterior quantiles cannot be used to form asymptotically valid confidence intervals in the sense of (15). However, typically in (25), $(\Lambda_{n,\theta_0}, \sqrt{n} g_n(\theta_0)) \rightsquigarrow \mathcal{Z} = \{\mathcal{Z}_\Delta, \mathcal{Z}_{g_n}\}$, which can be consistently estimated by some $\hat{\mathcal{Z}} = \{\hat{\mathcal{Z}}_\Delta, \hat{\mathcal{Z}}_{g_n}\} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \mathcal{Z}$. Then using any $\hat{R} \xrightarrow{p} R$, $\hat{H} \xrightarrow{p} H_0$ and $\hat{G} \xrightarrow{p} G_0$, (25) can be consistently estimated by

$$\hat{R}(\hat{R}'\hat{H}\hat{R})^{-1}\hat{R}'\hat{\mathcal{Z}}_\Delta - \left(I - \hat{R}(\hat{R}'\hat{H}\hat{R})^{-1}\hat{R}'\hat{H}\right)\hat{G}(\hat{G}'\hat{G})^{-1}\hat{\mathcal{Z}}_{g_S}.$$

Analogs of Theorem 4 and result (14) of Theorem 5 can also be developed with estimated or simulated constraints.

4.2 Constrained Method of Moments

Let $\ell_n(\theta) = \ell(\theta) + o_P(1) \in \mathbb{R}^{d_\ell}$, $g_n(\theta) = g(\theta) + o_P(1) \in \mathbb{R}^{d_g}$ be two sets of sample moment conditions. Instead of weighting all moment conditions using the inverted covariance matrix, one might wish to enforce $g_n(\theta) = 0$ while applying sample weights $\hat{W}_l = W_l + o_P(1)$ to $\ell_n(\theta)$. This is a special case of section 4.1 and (25) with $\hat{Q}_n(\theta) = \ell_n(\theta)' \hat{W}_l \ell_n(\theta)$,

$$Q(\theta) = \ell(\theta)' W_l \ell(\theta), \quad \Delta_{n,\theta_0} = L(\theta_0) W_l \sqrt{n} \ell_n(\theta_0), \quad \text{and} \quad L(\theta) = \frac{\partial}{\partial \theta'} \ell(\theta).$$

An alternative to section 4.1 is combining ℓ_2 with a GMM objective function, and defining $\bar{\theta}_S$ as $\bar{Q}_n(\bar{\theta}_S) \geqslant \sup_{\theta \in \Theta} \bar{Q}_n(\theta) - o_P(n^{-1})$, where for $\hat{W} = \operatorname{diag}(\hat{W}_l, \lambda_n \hat{W}_g)$, $\hat{W}_g = W_g + o_P(1)$,

$$\bar{Q}_n(\theta) = -\ell_n(\theta)' \hat{W}_l \ell_n(\theta) - \lambda_n g_n(\theta)' \hat{W}_g g_n(\theta) = -\begin{pmatrix} \ell_n(\theta)' & g_n(\theta)' \end{pmatrix} \hat{W} \begin{pmatrix} \ell_n(\theta)' & g_n(\theta)' \end{pmatrix}'.$$

Without optimal weighting, the information matrix equality does not hold in this model, but the usual sandwich variance estimate will continue to provide consistent inference.

Denote $L_0 = L(\theta_0)$, and $H_0 = L_0 W_l^0 L_0'$. Using $\hat{L} \xrightarrow{p} L_0$, $\hat{G} \xrightarrow{p} G_0$, and $\hat{H} \xrightarrow{p} H_0$, the sandwich formula approximates (25) by, for $\hat{M} = \begin{pmatrix} \hat{L} & \hat{G} \end{pmatrix}$, $\hat{H} = \hat{L} \hat{W}_l \hat{L}'$, $\hat{\Delta} = \hat{L} \hat{W}_l \hat{\mathcal{Z}}_l$,

$$\left(\hat{M}' \hat{W} \hat{M}\right)^{-1} \left(\hat{L} \hat{W}_l \hat{\mathcal{Z}}_l + \lambda_n \hat{G} \hat{W}_g \hat{\mathcal{Z}}_g\right) = \left(\hat{H} + \lambda_n \hat{G} \hat{W}_g \hat{G}'\right)^{-1} \left(\hat{\Delta} + \lambda_n \hat{G} \hat{W}_g \hat{\mathcal{Z}}_g\right),$$

where $(\hat{\mathcal{Z}}_l, \hat{\mathcal{Z}}_g) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} N(0, \Omega)$, $\Omega = \operatorname{AsyVar}(\sqrt{n}\ell_n(\theta_0), \sqrt{n} g_n(\theta_0))$. Let $\lambda_n \to \infty$ and the parameters be identified: $\operatorname{rank}(L_0 \; G_0) = \dim(\theta)$. There are two cases to consider. In case (1), $\operatorname{rank}(G_0) = \dim(\theta)$ and $\hat{G} \hat{W}_g \hat{G}'$ is invertible, then $\ell_n(\theta)$ are asymptotically negligible, and $\hat{W}_g$ can be optimally chosen for valid posterior inference, since

$$\begin{aligned} &\left(\hat{H} + \lambda_n \hat{G} \hat{W}_g \hat{G}'\right)^{-1} \left(\hat{\Delta} + \lambda_n \hat{G} \hat{W}_g \hat{\mathcal{Z}}_g\right) \\ &= \left(\lambda_n^{-1} \hat{H} + \hat{G} \hat{W}_g \hat{G}'\right)^{-1} \left(\lambda_n^{-1} \hat{\Delta} + \hat{G} \hat{W}_g \hat{\mathcal{Z}}_g\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \left(\hat{G} \hat{W}_g \hat{G}'\right)^{-1} \hat{G} \hat{W}_g \hat{\mathcal{Z}}_g. \end{aligned}$$

In case (2), if $\operatorname{rank}(G_0) < \dim(\theta)$, then (25) will apply:

$$\left(\hat{H} + \lambda_n \hat{G}\hat{W}_g\hat{G}'\right)^{-1}\left(\hat{\Delta} + \lambda_n \hat{G}\hat{W}_g\hat{\mathcal{Z}}_g\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} (25).$$

To show this, let $J = \operatorname{diag}\left(\lambda_n^{-1} I_{d_g}, I_{d_l}\right)$ and $\hat{B} = \left(\hat{G}, \hat{R}\right)'$, and manipulate the LHS as

$$\begin{aligned}
&\hat{B}'\left(J\left(\hat{B}\hat{H}\hat{B}' + \lambda_n \hat{B}\hat{G}\hat{W}_g\hat{G}'\hat{B}'\right)\right)^{-1} J\hat{B}\left(\hat{\Delta} + \lambda_n \hat{G}\hat{W}_g\hat{\mathcal{Z}}_g\right) \\
&\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} B'\begin{pmatrix} (G'G)\, W_g\, (G'G) & 0 \\ R'HG & R'HR \end{pmatrix}^{-1}\begin{pmatrix} (G'G)\, W_g\hat{\mathcal{Z}}_g \\ R'\hat{\Delta} \end{pmatrix}.
\end{aligned}$$

Completing the calculation shows that this indeed does not depend on W_g:

$$R\left(R'HR\right)^{-1} R'\hat{\Delta} - \left(I - R\left(R'HR\right)^{-1} R'H\right) G_0\left(G_0'G_0\right)^{-1}\hat{\mathcal{Z}}_g \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} (25).$$

Remark 7. The bootstrap (multinomial or wild) can also be used for inference: $(\hat{\mathcal{Z}}_l, \hat{\mathcal{Z}}_g) = \sqrt{n}(\ell_n^*(\hat{\theta}), g_n^*(\theta))$, where for example, $g_n^*(\hat{\theta}) = g_n(X_n^*, \hat{\theta})$, $\ell_n^*(\hat{\theta}) = \ell_n(X_n^*, \hat{\theta})$, or $\ell_n^*(\hat{\theta}) = n^{-1}\sum_{i=1}^n \xi_i^* \ell(X_i, \hat{\theta})$, for i.i.d. $\xi_i^* > 0$, $E\xi_i^* = 1$.

Remark 8. It is often of empirical interest to conduct inference on a function of the parameter and the data. Suppose $\hat{Q}_n(\theta) = \hat{Q}_n(\theta_1)$ and $\theta_{2,0} = \eta(\theta_{1,0})$ is the policy function of interest, which is estimated by $\hat{\theta}_2 = \eta(\mathcal{X}_n, \hat{\theta}_1)$. Setting $g_n(\theta) = \theta_2 - \eta(\mathcal{X}_n, \theta_1)$ in the present framework enables inference for $\sqrt{n}(\hat{\theta}_2 - \theta_{2,0})$ as part of $\sqrt{n}(\hat{\theta} - \theta_0)$.

Remark 9. When the constraints are estimated with noise, statistically a more efficient estimator can be obtained by not enforcing the estimated constraints but by instead stacking up the estimated constraints with the other sample moment conditions implied by the model. A joint generalized method of moment estimator can be obtained by using the estimated joint variance-covariance matrix of $(\ell_n(\theta_0), g_n(\theta_0))$. The choice of enforcing $g_n(\theta_0)$ instead of weighting it according to its sample variation needs to be based on a priori reasoning that is beyond merely achieving statistical efficiency. In the special case when $(\ell_n(\theta), g_n(\theta))$ jointly exactly identifies θ, there is no difference between optimally weighting and strictly

enforcing the moments. The case with known constraints $g(\theta)$ is also a special case of section 4.1, 4.2, and optimally weighted GMM when $Var(\sqrt{n}g_n(\theta)) = o(1)$.

4.3 Lagrange Multiplier Representation

An alternative to Theorems 2, 3 and 8 is by means of a Lagrange multiplier representation of the estimator. The Lagrange multiplier representation has the form (Silvey (1975), Sections 3.10.2, 4.7.3)

$$\begin{bmatrix} -H_0 + G_0 G_0' & G_0 \\ G_0' & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\bar{\theta} - \theta_0) \\ \lambda \end{bmatrix} = \begin{bmatrix} -\Delta_{n,\theta_0} \\ 0 \end{bmatrix} \tag{26}$$

where H_0 and G_0 are as in Theorem 3. The matrix on the left hand side is non-singular (Silvey (1975), Appendix A).

It can be verified that equation (8) asymptotically satisfies (26), together with λ defined as

$$\lambda = -\left(G_0' G_0\right)^{-1} G_0' \left(I - H_0 R \left(R' H_0 R\right)^{-1} R'\right) \Delta_{n,\theta_0}.$$

The second set of rows in (26), hold by definition of (8) and that $G_0' R = 0$. The first set of rows of (26) can be written as $G_0 \lambda = -\left(I - H_0 R \left(R' H_0 R\right)^{-1} R'\right) \Delta_{n,\theta_0}$. Or by the definition of λ, $G_0 \left(G_0' G_0\right)^{-1} G_0' \left(I - H_0 R \left(R' H_0 R\right)^{-1} R'\right) = \left(I - H_0 R \left(R' H_0 R\right)^{-1} R'\right)$. This can be verified to hold by replacing $G_0 \left(G_0' G_0\right)^{-1} G_0' = I - R \left(R' R\right)^{-1} R'$.

When the constraints are estimated in Theorem 8, (26) is replaced by

$$\begin{bmatrix} -H_0 + G_0 G_0' & G_0 \\ G_0' & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\bar{\theta} - \theta_0) \\ \lambda \end{bmatrix} = \begin{bmatrix} -\Delta_{n,\theta_0} \\ -\sqrt{n} g_n(\theta_0) \end{bmatrix}. \tag{27}$$

The second equation is now replaced by $G_0' \sqrt{n} \left(\bar{\theta} - \theta_0\right) = -\sqrt{n} g_n(\theta_0)$, which corresponds to a mean value expansion of the constraints $\sqrt{n} g_n\left(\bar{\theta}\right) = 0$. It has also be shown that (27) is

asymptotically satisfied by (25) together with

$$\begin{aligned}\lambda = &-\left(G_0'G_0\right)^{-1}G_0'\left(I - H_0R\left(R'H_0R\right)^{-1}R'\right)\Delta_{n,\theta_0} \\ &-\left(G_0'G_0\right)^{-1}G_0'\left(H_0 - G_0G_0'\right)\left(I - R\left(R'H_0R\right)^{-1}R'H_0\right)G_0\left(G_0'G_0\right)^{-1}\sqrt{n}g_n\left(\theta_0\right).\end{aligned} \tag{28}$$

4.4 One-Step Iteration

With an ℓ_2 penalty the one step iteration methods in Gallant (1987) and Robinson (1988) apply. Let $\tilde{\theta} = \theta_0 + O_P\left(\frac{1}{\sqrt{n}}\right)$ be an initial $\sqrt{n}$ consistent estimate of θ_0, and consider the following iteration:

$$\theta^- = \tilde{\theta} - \left(\tilde{H} + \lambda_n\tilde{G}\tilde{G}'\right)^{-1}\left(\tilde{\nabla} + \lambda_n\tilde{G}g_n\left(\tilde{\theta}\right)\right),$$

where for Gauss-Newton iteration,

$$\tilde{L} = \frac{\partial}{\partial\theta'}\ell_n\left(\tilde{\theta}\right), \quad \tilde{H} = \tilde{L}\hat{W}_l\tilde{L}', \quad \tilde{\nabla} = \tilde{L}\hat{W}_l\ell_n\left(\tilde{\theta}\right).$$

Using Taylor expansions and calculations similar to the previous section we will show that the one step estimator has the following influence function representation.

Theorem 9. *Under Assumptions 1–4 and 11,* $\sqrt{n}\left(\theta^- - \theta_0\right) = (25) + o_P\left(1\right).$

5 Monte Carlo

We investigate the empirical coverage frequencies of our posterior quantile intervals using the instrumental variable quantile regression example in Chernozhukov and Hong (2003). Our data are generated according to

$$Y = \alpha_0 + D'\beta_0 + u, u = \frac{1}{5}\left(1 + \sum_{i=1}^{3}D_{(i)}\right)\epsilon, D \sim expN(0, I_3), \epsilon \sim N(0,1)$$

Define $\theta \equiv \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ and $Z = \begin{pmatrix} 1 & D \end{pmatrix}$. The objective function is the GMM objective:

$$\hat{Q}_n(\theta) = -\frac{1}{2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta)\right)' W_n(\theta)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta)\right)$$

$$\frac{1}{n}\sum_{i=1}^{n} m_i(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2} - 1\left(y_i \leqslant \alpha + D_i'\beta\right)\right) z_i$$

$$W_n(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n} m_i(\theta)' m_i(\theta)\right)^{-1}$$

We use an adaptive Laplace prior of the form

$$\pi(\theta) = \prod_{j=1}^{4}\left(\frac{\lambda_n \hat{w}_j}{2}\right) e^{-\lambda_n \hat{w}_j |\theta_j|}$$

The adaptive weights are $\hat{w}_j = \frac{1}{\left|\hat{\theta}_j^{OLS}\right|^{\gamma}}$, for some $\gamma \geqslant 0$. We fix a particular value of λ_n that satisfies $\frac{\lambda_n}{\sqrt{n}} \to 0$, $\lambda_n n^{\gamma/2-1} \to 0$, and $\lambda_n n^{(\gamma-1)/2} \to \infty$ and use the Metropolis Hastings algorithm to construct our Markov chain for θ. The Metropolis Hastings Sampler has the following steps:

1. Initialize $\theta^{(0)} = (Z'Z)^{-1}(Z'Y)$.

2. For periods b =1 to B:

   (a) For parameters $j = 1$ to 4:

      i. Draw $\xi_j = \theta_j^{(b-1)} + N\left(0, \sigma_j^2\right)$

      ii. $\theta_j^{(b)} = \begin{cases} \xi_j & \text{wp } \rho\left(\theta_j^{(b-1)}, \xi_j\right) \\ \theta_j^{(b-1)} & \text{wp } 1 - \rho\left(\theta_j^{(b-1)}, \xi_j\right) \end{cases}$ where $\rho\left(\theta_j^{(b-1)}, \xi_j\right) = min\left(\frac{exp(\hat{Q}_n(\xi_j))\pi(\xi_j)}{exp(\hat{Q}_n(\theta_j^{(b-1)}))\pi(\theta_j^{(b-1)})}, 1\right)$

The standard deviations σ_j of the transition kernel are initialized to 0.1 and then adjusted every 100 periods to maintain an acceptance rate of approximately 50%. After achieving the desired acceptance rate, we grow another chain for B periods while keeping σ_j fixed. The quantiles θ_α^* of the distribution of θ's drawn from this chain are used to form confidence intervals.

Table 1 shows the empirical coverage of the equal-tailed and symmetric posterior quantile intervals averaged across $R = 1000$ simulations. The equal-tailed interval is given by $\left(\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2}\right)$, and the symmetric interval is given by $\left(\theta^*_{1/2} - c_{1-\alpha}, \theta^*_{1/2} + c_{1-\alpha}\right)$ where $c_{1-\alpha}$ is the $1-\alpha$ percentile of the absolute value of the demedianed posterior draws.

The nominal level is 95%. We fix $\alpha_0 = 1$, $\beta_0' = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ and use $\lambda_n = n^{1/4}$ and $\gamma = 1$. We grow the chain for $B = 30np$ periods and burn in the first $t = 5np$ periods. We also report the normalized posterior chain average standard deviation $\frac{1}{R}\sum_{i=1}^{R} \sqrt{n}\left(\theta^*_{(1-\alpha/2)}(i) - \theta^*_{(\alpha/2)}(i)\right) / \left(z_{1-\alpha/2} - z_{\alpha/2}\right)$.

Table 1: Adaptive IV Quantile

	Equal-tailed	Symmetric	Standard Dev
α_0	0.949	0.952	1.204
β_{01}	0.999	0.999	0.285
β_{02}	0.999	0.999	0.301
β_{03}	0.998	0.999	0.311

$n = 2000$, $\lambda_n = n^{1/4}$, 1000 simulations.

Notice that the nonzero coefficient has close to 95% coverage while the zero coefficients have close to 100% coverage. This is due to the quasi-posterior's faster rate of contraction along the correctly specified constraints than along the misspecified constraints (see Theorem 7).

To investigate the empirical coverage of the constrained IV Quantile Regression estimator, we consider a data generating process of the form

$$Y = \theta_{0,1} + X_2\theta_{0,2} + X_3\theta_{0,3} + u, u = \frac{1}{5}\left(1 + \sum_{i=1}^{3} D_{(i)}\right)\epsilon$$

where $D \sim expN(0, I_3)$, $\epsilon \sim N(0,1)$, $X = \begin{pmatrix} X_2 & X_3 \end{pmatrix} \sim N(0, \Omega)$, $\Omega \equiv \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$. The true $\theta_0 = [0.5, 0.5, 0.5]'$.

Define $Z = \begin{pmatrix} 1 & X \end{pmatrix}$. Our moment conditions are given by

$$\frac{1}{n}\sum_{i=1}^{n} m_i(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} \left(0.5 - 1\left(y_i \leqslant \theta_1 + \theta_2 x_{2i} + \theta_3 x_{3i}\right)\right) z_i$$

And the objective function and weighting matrix are given by

$$\hat{Q}_n(\theta) = -\frac{1}{2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta)\right)' W_n(\theta) \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta)\right)$$
$$W_n(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n} m_i(\theta)' m_i(\theta)\right)^{-1}$$

We first use an adaptive Laplace prior of the form

$$\pi(\theta) = \prod_{j=1}^{3}\left(\frac{\lambda_n \hat{w}_j}{2}\right) e^{-\lambda_n \hat{w}_j |\theta_j|}$$

where $\hat{w}_j = \frac{1}{\left|\hat{\theta}_j^{OLS}\right|^{\gamma}}$. The empirical coverage frequencies for the equal-tailed and symmetric intervals are shown in Table 2. We use $n = 2000$,$\alpha = 0.05$, $R = 2000$, $B = 30np$, and the number of burn-in periods is $t = 5np$. Notice that the empirical coverage is close to the nominal level of 95% for all three parameters, and the equal-tailed interval gives slightly higher coverage than the symmetric intervals.

Table 2: Adaptive IV Quantile

	Equal-tailed	Symmetric	Standard Dev
θ_1	0.944	0.942	1.120
θ_2	0.949	0.948	1.208
θ_3	0.957	0.954	1.228

$n = 2000$, $\lambda_n = n^{1/4}$, $\gamma = 1$, 2000 simulations.

Now we would like to nonadaptively impose a nonlinear constraint of the form $\theta_2^2 + sin(\theta_3) = \frac{1}{4} + sin\left(\frac{1}{2}\right)$. Our prior then becomes

$$\pi(\theta) \propto e^{-\lambda_n \sqrt{n} |g(\theta)|}$$

where $g(\theta) = \theta_2^2 + sin(\theta_3) - \left(\frac{1}{4} + sin\left(\frac{1}{2}\right)\right)$. Table 3 shows the empirical coverage frequencies and normalized posterior chain average standard deviation for $\lambda_n = n^{1/4}$, $n = 2000$, $R = 2000$, and $t = 5np$. Both the equal-tailed and symmetric intervals give coverage close to the nominal level of 95%.

Table 3: IV Quantile with Correctly Specified Constraint

	Equal-tailed	Symmetric	Standard Dev
θ_1	0.952	0.949	1.131
θ_2	0.968	0.966	0.930
θ_3	0.968	0.965	1.052

$n = 2000$, $\lambda_n = n^{1/4}$, 2000 simulations, $g(\theta) = \theta_2^2 + sin(\theta_3) = \frac{1}{4} + sin\left(\frac{1}{2}\right)$.

As a point of comparison, we also consider a model without the constraint, which effectively sets $\pi(\theta) = 0$. As seen in Table 4, there is some slight over coverage, and the standard deviation is higher. This is to be expected since imposing the constraint reduces the size of the search space for the MCMC routine.

Table 4: IV Quantile without Constraint

	Equal-tailed	Symmetric	Standard Dev
θ_1	0.952	0.950	1.126
θ_2	0.957	0.956	1.214
θ_3	0.972	0.971	1.236

$n = 2000$, 2000 simulations.

Now suppose that our constraint is misspecified: $g(\theta) = 0.5$. If we continue to use a nonadaptive prior, the asymptotic bias diverges due to the nonvanishing penalty term, which results in severe undercoverage, as shown in Table 5.

Suppose we instead use an adaptive prior:

$$\pi(\theta) \propto e^{-\frac{\lambda_n}{|g(\hat{\theta}_{OLS})-0.5|^{\gamma}}|g(\theta)-0.5|}$$

Table 6 shows that the empirical coverage frequencies of the equal-tailed and symmetric intervals are close to the nominal level of 95%.

Table 5: Nonadaptive IV Quantile with Misspecified Constraint

	Equal-tailed	Symmetric	Standard Dev
θ_1	0.826	0.812	1.398
θ_2	0	0	0.841
θ_3	0.381	0.407	1.468

$n = 2000$, 2000 simulations, $g(\theta) = 0.5$.

Table 6: Adaptive IV Quantile with Misspecified Constraint

	Equal-tailed	Symmetric	Standard Dev
θ_1	0.949	0.950	1.127
θ_2	0.946	0.944	1.227
θ_3	0.970	0.965	1.234

$n = 2000$, $\lambda_n = n^{1/4}$, 2000 simulations, $g(\theta) = 0.5$.

6 Example: Conditional Moment Restrictions

We consider estimation of a conditional density

$$f(x_t \,|\, x_{t-1}, \theta_{(1)}) \tag{29}$$

subject to conditional moment conditions

$$0 = \int m_j(x_t, x_{t-1}, \theta_{(2)})\, f(x_t \,|\, x_{t-1}, \theta_{(1)})\, dx_t \tag{30}$$

for all x_{t-1} and for $j = 1, \ldots, J$. The context is asset pricing for an endowment economy:

6.1 An Endowment Economy

Let C_t denote the annual consumption endowment. Let

$$R_{st} = (P_{st} + D_{st})/P_{s,t-1} \tag{31}$$

denote the gross return on an asset S that pays D_{st} per period and has price P_{st} at time t. Prices and payoffs are real.

The constant relative risk aversion utility function is

$$U = \sum_{t=0}^{\infty} \delta^t \left(\frac{C_t^{1-\gamma} - 1}{1-\gamma} \right) \tag{32}$$

where δ is the time preference parameter and γ is the coefficient of risk aversion (Lucus, 1978). The agent's intertemporal marginal rate of substitution is

$$\text{MRS}_{t-1,t} = \delta \left(\frac{C_t}{C_{t-1}} \right)^{-\gamma}. \tag{33}$$

The gross return on an asset S that pays D_{st} satisfies

$$1 = E_{t-1} \left(\text{MRS}_{t-1,t} R_{s,t} \right). \tag{34}$$

The following variables were constructed for the 86 years 1930 to 2015 as described in Subsection 6.2 below.

- s_t = log real gross stock return (value weighted NYSE/AMEX/NASDAQ).
- b_t = log real gross bond return (30 day T-bill return).
- c_t = log real per capita consumption growth (nondurables and services).

Let $x_t = (s_t, b_t, c_t)'$, $t = 1, \ldots, n$, denote these data. They are presumed to follow the trivariate model

$$f\left(x_t \,|\, x_{t-1}, \theta_{(1)}\right) = N(x_t \,|\mu_{t-1}, \Sigma_{t-1}) \tag{35}$$

with location parameter

$$\mu_{t-1} = b_0 + B x_{t-1} \tag{36}$$

and two different scale parameter specifications, namely, VAR

$$\Sigma_{t-1} = R_0 R_0' \tag{37}$$

and ARCH

$$\Sigma_{t-1} = R_0 R_0' + [\text{diag}(p_1, p_2, p_3)](x_{t-2} - \mu_{t-2})(x_{t-2} - \mu_{t-2})'[\text{diag}(p_1, p_2, p_3)]. \tag{38}$$

These densities require initial lags in estimation. We held out five lags so that the years 1930 to 1934 provide the initial lags and the years 1935 to 2015 provide the data for estimation.

Given the parameters $\theta_{(2)} = (\gamma, \delta)$ and x, one can compute the pricing errors

$$\begin{aligned} e_1(x_t, x_{t-1}, \theta_{(2)}) &= 1 - \exp(\text{mrs}_{t-1,t} + s_t) && (39) \\ e_2(x_t, x_{t-1}, \theta_{(2)}) &= 1 - \exp(\text{mrs}_{t-1,t} + b_t), && (40) \end{aligned}$$

where $\text{mrs}_{t-1,t} = \log(\text{MRS}_{t-1,t}) = \log \delta - \gamma c_t$. The pricing errors satisfy

$$0 = m_j(x_{t-1}, \theta_{(2)}) = \int e_j(x_t, x_{t-1}, \theta_{(2)})\, f(x_t \,|\, x_{t-1}, \theta_{(1)})\, dx_t \tag{41}$$

for $j = 1, 2$ and for all x_{t-1}. Equivalently, the pricing errors satisfy

$$0 = g_j(\theta) = \int \left[m_j(x_{t-1}, \theta_{(2)})\right]^2 f(x_{t-1}, \theta_{(1)})\, dx_{t-1} \tag{42}$$

for $j = 1, 2$, where $\theta = (\theta_{(1)}, \theta_{(2)})$, and $f(x, \theta_{(1)})$ is the stationary density implied by (35).

6.2 Data

The raw data for stock returns are value weighted returns including dividends for NYSE, AMEX, and NASDAQ from the Center for Research in Security Prices data at the Wharton Research Data Services web site (http://wrds.wharton.upenn.edu).

The raw data for returns on U.S. Treasury 30 day debt are from the Center for Research in Security Prices data at the Wharton Research Data Services web site.

The raw consumption data are personal consumption expenditures on nondurables and services obtained from Table 2.3.5 at the Bureau of Economic Analysis web site (http://www.bea.gov).

Raw data are converted from nominal to real using the annual consumer price index

obtained from Table 2.3.4 at the Bureau of Economic Analysis web site. Conversion of consumption to per capita is by means of the mid-year population data from Table 7.1 at the Bureau of Economic Analysis web site.

Simple statistics for these data are shown in the first panel of Table 3. They are plotted in Figure 1.

6.3 Implementation

The integral in (41) is computed by three dimensional Gaussian quadrature using a five point rule for each dimension (Golub and Welsch (1969)). The integral in (42) is computed by summing over the data, *viz*

$$g_j(\theta) = \sum_{t=6}^{n+1} \left[m_j(x_{t-1}, \theta_{(2)}) \right]^2 \tag{43}$$

Parameter estimates are computed using the ℓ_2 penalty with a non-adaptive Gaussian quasi-prior: $\kappa\left(\lambda_n \sqrt{n} g\left(\theta\right)\right) = \exp\left[-(n\lambda_n^2)(g_1^2(\theta) + {}_2^2\left(\theta\right))\right]$. $\lambda_n = 10^k$ was chosen by increasing k until plots of the estimated parameters stabilized with attention particularly focused on the stability of γ and δ. The optimal choice turned out to be $\lambda_n = 10^7$.

Standard errors are computed by the method in section 4.3. Note that Ω has zeroes as its last two rows and columns due to the fact that the parameters $\theta_{(2)}$ do not appear in $f\left(x_t \,|\, x_{t-1}, \theta_{(1)}\right)$ and hence Ω is singular. Similarly for the Hessian H_0. We estimate $-H_0$ and Ω using the average of the outer product of the scores of (35) evaluated at $\bar{\theta}$. We estimate G_0 by $\frac{\partial}{\partial \theta} g(\theta)$ evaluated at $\bar{\theta}$ with (43) used for $g(\theta)$.

6.4 Estimation Results

The moment conditions (42) are well known to be incompatible with U.S. data. Therefore, what is of interest in our analysis is how the law of motion (35) is distorted by imposing the conditions and how closely estimates align with partial equilibrium generalized method of moments (GMM) estimates. Tables 1 and 2 present estimates for the VAR and ARCH models, respectively for both unconstrained and constrained by (43). Simple statistics for simulations of $x_t = (s_t, b_t, c_t)$ of length 1000 are from these four estimated densities together with simple statistics for the data are shown in Table 3.

Table 1. VAR Maximum Likelihood Estimates

	Unconstrained		Constrained	
Parameter	Estimate	Std. Dev.	Estimate	Std. Dev.
$b_{0,1}$	0.12243	0.11929	-0.35499	0.16911
$b_{0,2}$	-0.02076	0.07092	0.02931	0.07226
$b_{0,3}$	0.05274	0.08378	-0.28397	0.17012
$B_{1,1}$	-0.10189	0.13523	0.03153	0.17053
$B_{2,1}$	0.13075	0.07228	0.21259	0.07579
$B_{3,1}$	0.43096	0.08202	0.11228	0.15240
$B_{1,2}$	-0.00097	0.13280	0.12268	0.19084
$b_{2,2}$	0.92498	0.05263	0.83099	0.08204
$B_{3,2}$	-0.01672	0.07313	0.43586	0.12731
$B_{1,3}$	-0.32357	0.11952	0.02162	0.19237
$B_{2,3}$	0.09940	0.07222	0.14567	0.07383
$B_{3,3}$	0.31255	0.09662	0.07672	0.19259
$R_{0,1,1}$	0.82012	0.08282	0.98122	0.13440
$R_{0,1,2}$	-0.00160	0.02591	0.00705	0.03570
$R_{0,2,2}$	0.36936	0.03005	0.38672	0.04490
$R_{0,1,3}$	0.06736	0.04896	0.12703	0.10800
$R_{0,2,3}$	-0.00216	0.03939	-0.09160	0.06211
$R_{0,3,3}$	0.56863	0.04459	0.89382	0.10618
γ			2.49375	2.43299
δ			0.99989	0.07807

Maximum likelihood estimates for the density (35) with location (36) and scale (37). (left two columns) and same subject to moment conditions (43) (right two columns). The data are as in Figure 1. For the constrained estimates $\lambda_n = 10^7$ and the 0%, 25%, 50%, 75%, 100% quantiles of the conditional moment conditions (41) evaluated at $\{x_{t-1}\}_{t=6}^{n+1}$ are -1.62e-4, -2.03e-5, 7.70e-6, 3.81e-5, 8.74e-5, respectively.

The primary distortion caused by imposing (43) occurs in the location parameters with little effect on scale parameters. While inspection of Tables 1 and 2 suggest this conclusion, it is readily apparent from inspection of Table 3.

Table 4 verifies that our estimates of the CRRA parameters γ and δ are in line with a partial equilibrium analysis. One might remark in passing that an assessment of the distortion of the law of motion cannot be obtained via a partial equilibrium analysis but can be with the methods proposed in this paper.

All estimates were computed using the Chernozhukov and Hong (2003) method with support conditions $\gamma > 0$ and $0 < \delta < 1$ and 500,000 repetitions after transients have died out. The modal value of these repetitions is the estimator reported in the tables.

Table 2. ARCH Maximum Likelihood Estimates

	Unconstrained		Constrained	
Parameter	Estimate	Std. Dev.	Estimate	Std. Dev.
$b_{0,1}$	0.11492	0.13225	-0.35825	0.15658
$b_{0,2}$	-0.05429	0.06187	-0.03955	0.06843
$b_{0,3}$	0.03479	0.08792	-0.24276	0.17066
$B_{1,1}$	-0.13287	0.16147	0.02529	0.14427
$B_{2,1}$	0.10029	0.06965	0.17427	0.07920
$B_{3,1}$	0.41579	0.08997	0.10771	0.13658
$B_{1,2}$	0.00263	0.14677	0.12602	0.17608
$b_{2,2}$	0.87614	0.05253	0.85400	0.08317
$B_{3,2}$	0.05831	0.06666	0.52984	0.12676
$B_{1,3}$	-0.32081	0.11486	0.01699	0.17895
$B_{2,3}$	0.06062	0.06961	0.11634	0.09260
$B_{3,3}$	0.21388	0.10752	0.07236	0.19473
$R_{0,1,1}$	0.81061	0.10965	0.92173	0.12753
$R_{0,1,2}$	-0.00014	0.02554	-0.01669	0.02900
$R_{0,2,2}$	0.28764	0.04132	0.31842	0.04763
$R_{0,1,3}$	0.03719	0.05189	0.10892	0.08646
$R_{0,2,3}$	0.07668	0.04901	-0.01727	0.07508
$R_{0,3,3}$	0.51085	0.07393	0.85212	0.13224
p_1	-0.15876	0.26488	0.00562	0.32288
p_2	-0.85521	0.19173	0.50124	0.20277
p_3	-0.43945	0.13356	0.17750	0.22826
γ			2.11113	1.82603
δ			0.99907	0.05057

Maximum likelihood estimates for the density (35) with location (36) and scale (38) (left two columns) and same subject to moment conditions (43) (right two columns). The data are as in Figure 1. For the constrained estimates $\lambda_n = 10^7$ and the 0%, 25%, 50%, 75%, 100% quantiles of the conditional moment conditions (41) evaluated at $\{x_{t-1}\}_{t=6}^{n+1}$ are -4.07e-4, -4.43e-5, -7.60e-7, 2.47e-5, 9.67e-5, respectively.

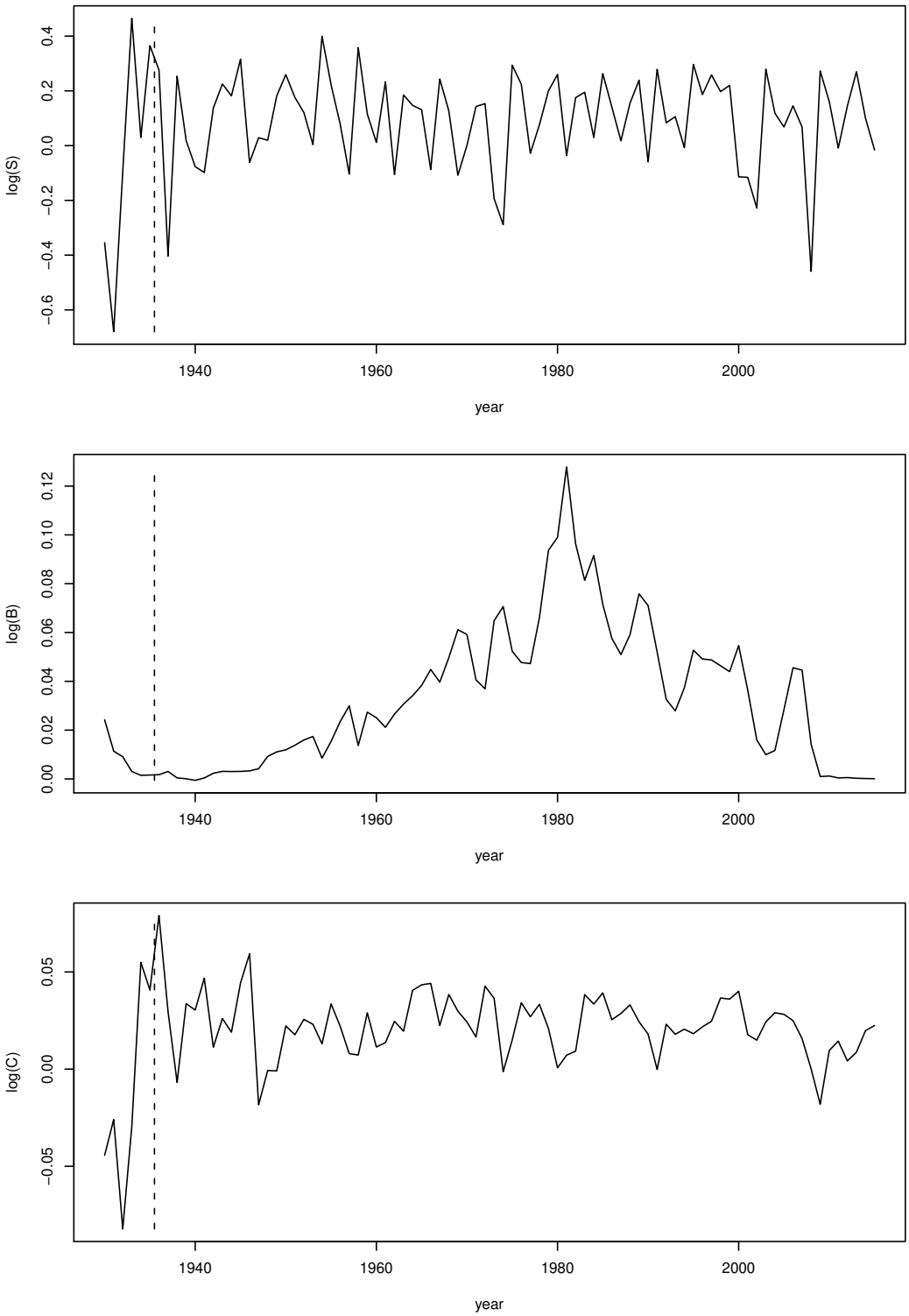

Figure 1. Annual Stock, Bond, and Consumption Data The data are real, annual, per capital consumption for the years 1930–2015 and real, annual gross stock and bond returns for the same years from BEA (2016) and CRSP (2016). The first five years are used to provide initial lags and are not otherwise used in estimation. These observations are to the left of the vertical dashed line. See Subsection 6.2 for complete details.

Table 3. Simple Statistics for the Data and Estimates

Series	Mean	Standard Deviation	Skewness	Excess Kurtosis
		Data		
s_t	0.09922	0.16802	-0.90325	0.90152
b_t	0.03348	0.02905	0.77644	0.13625
c_t	0.02276	0.01584	0.16411	1.35363
		Unconstrained VAR		
s_t	0.09395	0.17571	-0.01951	0.32526
b_t	0.03093	0.02763	-0.09774	-0.08085
c_t	0.02099	0.01682	-0.21009	1.31076
		Constrained VAR		
s_t	-0.03984	0.19318	0.02980	0.10306
b_t	-0.01682	0.03916	0.27078	-0.56675
c_t	-0.00473	0.02438	0.04412	0.07359
		Unconstrained ARCH		
s_t	0.09583	0.17469	-0.03754	0.29633
b_t	0.02408	0.03195	1.51407	10.4417
c_t	0.01994	0.01637	-0.24259	1.58554
		Constrained ARCH		
s_t	-0.06836	0.18096	0.00821	-0.01557
b_t	-0.04724	0.03887	0.30045	-0.56528
c_t	-0.02105	0.02560	0.07442	-0.00714

The first panel are simple statistics of annual data for the years 1935 through 2015. s_t is log real gross stock return. b_t is log real gross bond return. c_t is log real per capita consumption growth. The second panel are simple statistics computed for a simulation of length 1000 from the VAR density at parameter estimates shown as "Unconstrained" in Table 1. The third panel is the same as the second but evaluated at the "Constrained" estimates shown in Table 1. The fourth and fifth panels are the same as the third and fourth but for the ARCH estimates in Table 2.

Table 4. GMM Estimates

	Just Identified		Over Identified	
Parameter	Estimate	Std. Dev.	Estimate	Std. Dev.
γ	1.5708	303.73	1.9877	0.58828
δ	0.9999	6.9116	0.9999	0.01223

Generalized method of moments estimates (GMM). Data is as in Figure 1, denoted S_t, B_t, and C_t for gross stock returns, gross bond returns, and consumption growth, respectively. Just identified moments are $m_{1,t} = 1 - MRS_{t-1,t}S_t, m_{2,t} = 1 - MRS_{t-1,t}B_t$, where $MRS_{t-1,t}$ is given by (33). The additional, overidentifying moments are $m_{3,t} = \log(S_{t-1})m_{1,t}, m_{4,t} = \log(B_{t-1})m_{1,t}, m_{5,t} = \log(C_{t-1})m_{1,t}, m_{6,t} = \log(S_{t-1})m_{2,t}, m_{7,t} = \log(B_{t-1})m_{2,t}, m_{8,t} = \log(C_{t-1})m_{2,t}$.

7 Conclusion

This paper demonstrates the validity of using location functionals of the quasi-posterior distribution to perform inference on functions of parameters defined in terms of constrained optimization. We have considered the ℓ_1, ℓ_2, and ℓ_0 penalty functions and both nonadaptive and adaptive priors. The nonadaptive methods require constraints to be correctly specified, while adaptive methods provide valid inference even under misspecification of constraints. We also consider extensions allowing for simulated constraints and constraints that depend on the data.

References

Alhamzawi, Rahim, Keming Yu, and Dries F Benoit, "Bayesian adaptive Lasso quantile regression," *Statistical Modelling*, 2012, *12* (3), 279–297. 3

Amemiya, Takeshi, *Advanced Econometrics*, Harvard University Press, 1985. 12, 13, 47, 49, 52, 70

Anderson, Theodore Wilbur, *An introduction to multivariate statistical analysis*, Vol. 2, Wiley New York, 1958. 16

Belloni, Alexandre and Victor Chernozhukov, "On the computational complexity of MCMC-based estimators in large samples," *The Annals of Statistics*, 2009, pp. 2011–2055. 4

_ **and** _ , "ℓ_1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, 2011, *39* (1), 82–130. 6

_ **and** _ , "Posterior inference in curved exponential families under increasing dimensions," *The Econometrics Journal*, 2014, *17* (2), S75–S100. 15

Blackwell, David, *Approximate normality of large products*, Department of Statistics, University of California, 1985. 4

Chernozhukov, Victor and Han Hong, "A MCMC Approach to Classical Estimation," *Journal of Econometrics*, 2003, *115* (2), 293–346. 2, 4, 7, 10, 14, 15, 16, 18, 30, 39

Chib, Siddhartha, Minchul Shin, and Anna Simoni, "Bayesian estimation and comparison of moment condition models," *Journal of the American Statistical Association*, 2018, *113* (524), 1656–1668. 4

Florens, Jean-Pierre and Anna Simoni, "Gaussian processes and Bayesian moment estimation," *Journal of Business & Economic Statistics*, 2019, pp. 1–11. 4

Gallant, A Ronald, *Nonlinear statistical models*, John Wiley & Sons, 1987. 2, 5, 13, 30

_ , "Complementary bayesian method of moments strategies," *Journal of Applied Econometrics*, 2020. 9

Gallant, A. Ronald, "Nonparametric Bayes Subject to Overidentified Moment Conditions," *working paper*, 2020. 4, 8

Golub, G. H. and J. H. Welsch, "Calculation of Gaussian Quadrature Rules," *Mathematics of Computation*, 1969, *23*, 221–230. 38

Han, Aaron K, "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics*, 1987, *35* (2-3), 303–316. 5, 6

Hans, Chris, "Bayesian lasso regression," *Biometrika*, 2009, *96* (4), 835–845. 3

Hansen, Bruce E., "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 2016, *190*, 115–132. 2

Hausman, Jerry A and Tiemen Woutersen, "Estimating a semi-parametric duration model without specifying heterogeneity," *Journal of Econometrics*, 2014, *178*, 114–131. 6

Hong, Han and Jessie Li, "The numerical delta method," *Journal of Econometrics*, 2018, *206* (2), 379–394. 59, 61

_ **, Aprajit Mahajan, and Denis Nekipelov**, "Extremum estimation and numerical derivatives," *Journal of Econometrics*, 2015, *188* (1), 250–263. 72

Kitamura, Yuichi and Taisuke Otsu, "Bayesian analysis of moment condition models using nonparametric priors," *manuscript*, 2011. 4

Knight, K. and W. Fu, "Asymptotics for Lasso-Type Estimators," *Annals of Statistics*, 2000, *28* (5), 1356–1378. 24

Kosorok, Michael R, *Introduction to empirical processes and semiparametric inference*, Springer, 2007. 16

Leeb, Hannes and Benedikt M Pötscher, "Model selection and inference: Facts and fiction," *Econometric Theory*, 2005, *21* (01), 21–59. 4

_ **and** _ , "Guest editors' editorial: recent developments in model selection and related areas," *Econometric Theory*, 2008, *24* (02), 319–322. 4

_ **and** _ , "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, 2008, *142* (1), 201–211. 4

Leng, Chenlei, Minh-Ngoc Tran, and David Nott, "Bayesian adaptive lasso," *Annals of the Institute of Statistical Mathematics*, 2014, *66* (2), 221–244. 3

Newey, W. and D. McFadden, "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden, eds., *Handbook of Econometrics, Vol. 4*, North Holland, 1994, pp. 2113–2241. 12, 47, 50, 51

Park, T. and G. Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, 2008, *103* (482), 681–686. 2, 8

Politis, Dimitris N, Joseph P Romano, and Michael Wolf, *Subsampling*, Springer Science & Business Media, 1999. 59

Pollard, D., *Convergence of Stochastic Processes*, Springer Verlag, 1984. 72, 73

Pollard, David, "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 1991, *7* (2), 186–199. 49

Robinson, Peter M, "The stochastic difference between econometric statistics," *Econometrica: Journal of the Econometric Society*, 1988, pp. 531–548. 30

Schennach, Susanne M, "Bayesian exponentially tilted empirical likelihood," *Biometrika*, 2005, *92* (1), 31–46. 4

Sherman, Robert P., "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, 1993, *61*, 123–137. 5, 50, 52

Silvey, S. D., *Statistical Inference*, Chapman and Hall, London., 1975. 29

Tian, Lu, Jun S Liu, and LJ Wei, "Implementation of estimating function-based inference procedures with Markov chain monte carlo samplers," *Journal of the American Statistical Association*, 2007, *102* (479), 881–888. 4

Tibshirani, R., "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, *58* (1), 267–288. 6

van der Vaart, A., *Asymptotic Statistics*, Cambridge University Press, 2000. 14

van der Vaart, AW and Jon Wellner, *Weak Convergence and Empirical Processes*, Springer, 1996. 59, 61

Zhu, Ji, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie, "1-norm support vector machines," in "Advances in neural information processing systems" 2004, pp. 49–56. 6

Zou, Hui, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 2006, *101* (476), 1418–1429. 3, 4, 19, 21, 22

A Commonly used symbols

$h \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} \tilde{p}_\infty(\cdot)$	$\sup_x \left\| \int^x p_h(h \mid \mathcal{X}_n)\, dh - \int^x \tilde{p}_\infty(h)\, dh \right\| = o_P(1)$.
$\|f(\cdot) - g(\cdot)\|_\alpha$	$\int \|h\|^\alpha \|f(h) - g(h)\| dh$ for fixed $0 \leqslant \alpha < \infty$.
$\theta_g(\theta) = \arg\min_{\theta' \in \bar{\Theta}} \|\theta - \theta'\|$	projection of θ onto $\bar{\Theta}$.
$\bar{Q}_n(\theta) = \hat{Q}_n(\theta) - \frac{\lambda_n^p \sqrt{n}^p}{n} \sum_{j=1}^J \|g_j(\theta)\|^p$	non-adaptive penalized objective
$\bar{Q}_n(\theta) = \hat{Q}_n(\theta) - \frac{\lambda_n^p \sqrt{n}^p}{n} \sum_{m=1}^M \|\hat{w}_m g_m(\theta)\|^p$	adaptive penalized objective
$\hat{w}_m = \|g_m(\tilde{\theta})\|^{-\gamma}$ for some $\gamma > 0$	adaptive weights using $\tilde{\theta}$, a preliminary estimate of θ_0.
$b_n = \min(1, d_n)$	$d_n = \frac{\lambda_n^p \sqrt{n}^p}{n}$ (non-adaptive) or $d_n = \frac{\bar{\lambda}_n^p \sqrt{n}^p}{n}$ (adaptive)
$\bar{Q}_n^+(\theta) \equiv b_n^{-1} \bar{Q}_n(\theta)$	rescaled penalized objective
$Q^+(\theta) = b_n^{-1} Q(\theta) - b_n^{-1} \frac{\lambda_n^p \sqrt{n}^p}{n} \sum_{j=1}^J \|g_j(\theta)\|^p$	limit of rescaled non-adaptive penalized objective
$Q^+(\theta) = \bar{b}_n^{-1} Q(\theta) - \bar{b}_n^{-1} \frac{\lambda_n^p}{n} \sum_{m=1}^M \|\hat{w}_m g_m(\theta)\|^p$	limit of rescaled adaptive penalized objective
$B = (G_0, R)'$	where $R'G_0 = 0$ and $G_0 = \frac{\partial g}{\partial \theta}\big\|_{\theta=\theta_0}$
$D_n = \operatorname{diag}(\lambda_n I_J, I_{K-J})$	scaling matrix for non-adaptive posterior
$\bar{D}_n = \operatorname{diag}(\bar{\lambda}_n I_J, I_{K-J})$	scaling matrix for adaptive posterior

B Proofs

Example 6. We will make use of the following analytic example to illustrate how the relevant assumptions are employed following the proofs of Theorems 3 and 4 below:

$$\hat{Q}(\theta) = -\frac{1}{2}\left(\theta_2 - \bar{X}\right)^2, \quad g(\theta) = \theta_1 + \theta_2^2,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $X_i \sim i.i.d.(\theta_{20}, 1)$, $\theta_{1,0} = \theta_{2,0} = 0$. Therefore $Q(\theta) = \theta_2^2$. Both $\hat{Q}(\theta)$ and $g(\theta)$ are needed to identify and consistently estimate θ_1 and θ_2.

Proof of Theorem 1.

Part (a) concerns consistency of $\bar{\theta}$, which follows from standard arguments (e.g. Theorem 4.1.1 Amemiya (1985) and Theorem 2.1 Newey and McFadden (1994)).

Part (b) concerns consistency of θ^+. Recall that $\theta_g(\theta) = \arg\min_{\theta' \in \bar{\Theta}} \|\theta - \theta'\|$. By Assumption 2, for each $\delta > 0$, there exists $\epsilon_1(\delta) > 0$ such that $\|\theta - \theta_g(\theta)\| > \delta$ implies $|g_j(\theta)| > \epsilon_1(\delta)$ for all j. Likewise, by Assumption 4, for each $\delta > 0$, there exists $\epsilon_2(\delta) > 0$ such that

$\|\theta_0 - \theta_g(\theta)\| > \delta$ implies $Q\left(\theta_g\left(\theta\right)\right) - Q\left(\theta_0\right) < -\epsilon_2\left(\delta\right)$. We may choose $\epsilon_1(\cdot), \epsilon_2(\cdot)$ to be strictly increasing. Define $\epsilon(\delta) = \min(\epsilon_1(\delta), \epsilon_2(\delta))$ for all δ.

We first establish two preliminary results for later. Let $b_n = \min\left(1, d_n\right)$ for $d_n = \frac{\lambda_n^p \sqrt{n}^p}{n}$ and $\bar{Q}_n^+\left(\theta\right) \equiv b_n^{-1} \bar{Q}_n\left(\theta\right)$. By Assumptions 3 and 6(a),

$$\sup_{\theta \in \Theta} |\bar{Q}_n^+\left(\theta\right) - Q^+\left(\theta\right)| = o_P\left(1\right) \quad \text{for} \quad Q^+\left(\theta\right) = b_n^{-1} Q\left(\theta\right) - b_n^{-1} \frac{\lambda_n^p \sqrt{n}^p}{n} \sum_{j=1}^{J} |g_j\left(\theta\right)|^p. \tag{44}$$

For the second preliminary result, for any $\delta > 0$, we shall argue for the existence of $\eta\left(\delta\right) > 0$, such that

$$\|\theta^+ - \theta_0\| > \delta \quad \text{implies} \quad Q^+\left(\theta^+\right) < Q^+\left(\theta_0\right) - \eta\left(\delta\right). \tag{45}$$

On the event $\|\theta^+ - \theta_0\| > \delta$, for any $C > 0$, either (i) $\|\theta_g\left(\theta^+\right) - \theta_0\| > \left(1 - \frac{1}{C}\right)\delta$ and $\|\theta_g\left(\theta^+\right) - \theta^+\| < \frac{1}{C}\delta$, or (ii) $\|\theta_g\left(\theta^+\right) - \theta^+\| > \frac{1}{C}\delta$. By Assumption 4, C can be chosen sufficiently large such that in the first case (i),

$$\begin{aligned} Q\left(\theta^+\right) - Q\left(\theta_0\right) =& Q\left(\theta^+\right) - Q\left(\theta_g\left(\theta\right)\right) + Q\left(\theta_g\left(\theta\right)\right) - Q\left(\theta_0\right) \\ \leqslant& \epsilon\left(\frac{\delta}{C}\right) - \epsilon\left(\left(1 - \frac{1}{C}\right)\delta\right) \equiv -\eta\left(\delta\right). \end{aligned}$$

This implies that

$$Q^+\left(\theta^+\right) - Q^+\left(\theta_0\right) \leqslant b_n^{-1}\left(Q\left(\theta^+\right) - Q\left(\theta_0\right)\right) \leqslant -b_n^{-1}\eta\left(\delta\right) \leqslant -\eta\left(\delta\right).$$

In the second case (ii), $|g_j\left(\theta^+\right)| \geqslant \epsilon\left(\frac{1}{C}\delta\right) \equiv \bar{\epsilon}\left(\delta\right)$. Since $Q(\theta_0) \geqslant Q(\theta^+)$ by Assumption 4,

$$Q^+\left(\theta^+\right) - Q^+\left(\theta_0\right) \leqslant -b_n^{-1} d_n J \bar{\epsilon}\left(\delta\right)^p \leqslant -J\bar{\epsilon}\left(\delta\right)^p \equiv -\eta\left(\delta\right).$$

Finally, we have

$$\begin{aligned} P\left(\left\|\theta^{+}-\theta_{0}\right\|>\delta\right) &\leqslant P\left(Q^{+}\left(\theta^{+}\right)<Q^{+}\left(\theta_{0}\right)-\eta\left(\delta\right)\right) \\ &= P\left(Q^{+}\left(\theta^{+}\right)-\bar{Q}_{n}^{+}\left(\theta^{+}\right)<Q^{+}\left(\theta_{0}\right)-\bar{Q}_{n}^{+}(\theta_{0})+\bar{Q}_{n}^{+}(\theta_{0})-\bar{Q}_{n}^{+}\left(\theta^{+}\right)-\eta\left(\delta\right)\right) \\ &\leqslant P\left(\sup_{\theta\in\Theta}|\bar{Q}_{n}^{+}\left(\theta\right)-Q^{+}\left(\theta\right)|>\frac{\eta\left(\delta\right)}{3}\right)+P\left(\bar{Q}_{n}^{+}\left(\theta^{+}\right)-\bar{Q}_{n}^{+}\left(\theta_{0}\right)<-\frac{\eta\left(\delta\right)}{3}\right), \end{aligned}$$

where the first line follows from (45) and the last line from the law of total probability. The first term on the last line is $o(1)$ by (44), and the second is $o(1)$, since $\bar{Q}_{n}^{+}\left(\theta^{+}\right)\geqslant\bar{Q}_{n}^{+}\left(\theta_{0}\right)-o_{P}\left(1\right)$ by definition.

Lastly we prove part (c). Note that both ϕ_{τ}^{*} and ϕ^{*} can be written as special cases of M-estimators $\tilde{\phi}$ that minimize posterior loss of the form:

$$\tilde{\phi}=\arg\min_{\phi\in\phi(\Theta)}\Gamma_{n}\left(\phi\right)\equiv\frac{\int\rho\left(\phi-\phi\left(\theta\right)\right)\pi_{0}\left(\theta\right)e^{n\bar{Q}_{n}(\theta)}d\theta}{\int\pi_{0}\left(\theta\right)e^{n\bar{Q}_{n}(\theta)}d\theta},$$

where $\rho\left(u\right)$ is convex and Lipschitz with constant bounded on compact sets and is uniquely minimized at $u\equiv0$. By the Convexity Lemma in Pollard (1991) and Theorem 4.1.1 Amemiya (1985), it suffices to show that $\Gamma_{n}\left(\phi\right)=\Gamma\left(\phi\right)+o_{P}\left(1\right)$ for each fixed ϕ, where $\Gamma\left(\phi\right)=\rho\left(\phi-\phi\left(\theta_{0}\right)\right)$. For this purpose, note that for C denoting a generic constant,

$$\left|\Gamma_{n}\left(\phi\right)-\Gamma\left(\phi\right)\right|\leqslant C\frac{\int_{\Theta}\left\|\theta-\theta_{0}\right\|\pi_{0}\left(\theta\right)e^{n\bar{Q}_{n}(\theta)}d\theta}{\int_{\Theta}\pi_{0}\left(\theta\right)e^{n\bar{Q}_{n}(\theta)}d\theta}\leqslant C\delta+C\frac{\int_{\left\|\theta-\theta_{0}\right\|>\delta}\pi_{0}\left(\theta\right)e^{n\left(\bar{Q}_{n}(\theta)-\bar{Q}_{n}(\theta_{0})\right)}d\theta}{\int_{\Theta}\pi_{0}\left(\theta\right)e^{n\left(\bar{Q}_{n}(\theta)-\bar{Q}_{n}(\theta_{0})\right)}d\theta}$$

It then suffices for the last term to be $o_{P}\left(1\right)$. Separately, we bound, w.p.a. 1, $nb_{n}\to\infty$,

$$\begin{aligned} (A) &= \int_{\left\|\theta-\theta_{0}\right\|>\delta}\pi_{0}\left(\theta\right)e^{n\left(\bar{Q}_{n}(\theta)-\bar{Q}_{n}(\theta_{0})\right)}d\theta=\int_{\left\|\theta-\theta_{0}\right\|>\delta}\pi_{0}\left(\theta\right)e^{nb_{n}\left(\bar{Q}_{n}^{+}(\theta)-\bar{Q}_{n}^{+}(\theta_{0})\right)}d\theta \\ &\leqslant Ce^{nb_{n}(-\eta(\delta)+o_{P}(1))}\leqslant Ce^{-nb_{n}\eta(\delta)/2}. \end{aligned}$$

For the denominator, there are three cases. First suppose d_{n} converges to a positive constant. Applying (44) twice and using the fact that $Q\left(\theta\right)$ and $g\left(\theta\right)$ have bounded derivatives,

$$\sup_{\left\|\theta-\theta_{0}\right\|\leqslant\Delta_{1}}|\bar{Q}_{n}^{+}\left(\theta\right)-\bar{Q}_{n}^{+}\left(\theta_{0}\right)|\leqslant\Delta_{2}+o_{P}\left(1\right)$$

for some Δ_1, Δ_2 sufficiently small. Then w.p.a. 1,

$$(B) = \int_\Theta \pi_0(\theta)\, e^{n\left(\bar{Q}_n(\theta)-\bar{Q}_n(\theta_0)\right)} d\theta \geqslant \int_{\|\theta-\theta_0\|<\Delta_1} \pi_0(\theta)\, e^{nb_n(-\Delta_2+o_P(1))} \geqslant c\Delta_1^K e^{-nb_n\Delta_2/2}.$$

Therefore w.p.a. 1, for Δ_2 sufficiently smaller than η,

$$(A)/(B) \leqslant C\Delta_1^{-d} e^{-nb_n(\eta/2-\Delta_2/2)} \longrightarrow 0.$$

Second, suppose $d_n \to 0$. Then, similar to the first case, we can find Δ_1, Δ_2 sufficiently small such that

$$\begin{aligned} &\sup_{\|\theta-\theta_0\|\leqslant\Delta_1 d_n} |\bar{Q}_n^+(\theta) - \bar{Q}_n^+(\theta_0)| \leqslant \Delta_2 + o_P(1) \\ \Longrightarrow \quad &(B) \geqslant \int_{\|\theta-\theta_0\|<\Delta_1 d_n} \pi_0(\theta)\, e^{nb_n(-\Delta_2+o_P(1))} d\theta \geqslant c\Delta_1^K d_n^K e^{-nb_n\Delta_2/2}. \end{aligned}$$

By Assumption 6(a), w.p.a. 1, for Δ_2 sufficiently smaller than η,

$$(A)/(B) \leqslant C\Delta_1^{-K} d_n^{-K} e^{-nb_n(\eta/2-\Delta_2/2)} \longrightarrow 0.$$

Finally, suppose $d_n \to \infty$. Then we can find Δ_1, Δ_2 small enough such that

$$\begin{aligned} &\sup_{\|\theta-\theta_0\|\leqslant\Delta_1 d_n^{-1/p}} |\bar{Q}_n^+(\theta) - \bar{Q}_n^+(\theta_0)| \leqslant \Delta_2 + o_P(1) \\ \Longrightarrow\ &(B) \geqslant \int_{\|\theta-\theta_0\|<\Delta_1 d_n^{-1/p}} \pi_0(\theta)\, e^{nb_n(-\Delta_2+o_P(1))} d\theta \geqslant c\Delta_1^K d_n^{K/p} e^{-nb_n\Delta_2/2}. \end{aligned}$$

Therefore, w.p.a. 1, $(A)/(B) \leqslant C\Delta_1^{-K} d_n^{-K/p} e^{-nb_n(\eta/2-\Delta_2/2)} \longrightarrow 0$. ∎

Proof of Theorem 2.

Consider first the case when H_0 is nonsingular. Let $h = \sqrt{n}(\theta - \theta_0)$. Note that Theorem 1 in Sherman (1993) goes through verbatim under constraints, since its condition (i) holds when H_0 is nonsingular. Hence $\bar{h} = \sqrt{n}\left(\bar{\theta} - \theta_0\right) = O_P(1)$.

Define $h^* = \arg\max_{h:g(\theta_0+h/\sqrt{n})=0} \Delta_{n,\theta_0}' h - \frac{1}{2} h' H_0 h$, and $h^+ = R\left(R' H_0 R\right)^{-1} R' \Delta_{n,\theta_0}$. By the same arguments as those following Theorem 9.1 of NM 1994, $h^* = h^+ + o_P(1)$. Next

Taylor expanding $g\left(\theta_0+\bar{h}/\sqrt{n}\right)=0$ shows that $\left(G_0+O_P\left(\frac{1}{\sqrt{n}}\right)\right)'\bar{h}=0$, implying $G_0'\bar{h}=O_P\left(\frac{1}{\sqrt{n}}\right)$ since $\bar{h}=O_P(1)$. Also $G_0'h^+=0$.

By definition of $\bar{h}$, for $\Gamma_n(h)=n\left(\hat{Q}_n\left(\theta_0+\frac{h}{\sqrt{n}}\right)-\hat{Q}_n(\theta_0)\right)$, $\Gamma_n\left(\bar{h}\right)\geqslant\Gamma_n(h^*)-o_P(1)$. Invoking Assumption 5 on both sides,

$$\Delta_{n,\theta_0}'\bar{h}-\frac{1}{2}\bar{h}'H_0\bar{h}\geqslant\Delta_{n,\theta_0}'h^*-\frac{1}{2}h^{*'}H_0h^*-o_P(1)=\Delta_{n,\theta_0}'h^+-\frac{1}{2}h^{+'}H_0h^+-o_P(1) \tag{46}$$

If we write $\bar{\eta}=\bar{h}-h^+$, then this can be rewritten as

$$\frac{1}{2}\bar{\eta}'H_0\bar{\eta}-\Delta_{n,\theta_0}'\bar{\eta}+\bar{\eta}'H_0h^+\leqslant o_P(1)$$

Let $B=(G_0,R)'$, $\bar{v}=(\bar{v}_1'\ \ \bar{v}_2')'=B\bar{\eta}$, $\bar{\eta}=B^{-1}\bar{v}$, $B^{-1}=\left[G_0\left(G_0'G_0\right)^{-1},R\left(R'R\right)^{-1}\right]$. Then

$$\frac{1}{2}\bar{v}'B^{-1'}H_0B^{-1}\bar{v}-\Delta_{n,\theta_0}'B^{-1}\bar{v}+\bar{v}'B^{-1'}H_0h^+\leqslant o_P(1). \tag{47}$$

Since $\bar{v}_1=G_0'\left(\bar{h}-h^+\right)=O_P\left(\frac{1}{\sqrt{n}}\right)=o_P(1)$, this translates into

$$\frac{1}{2}\bar{v}_2'\left(R'R\right)^{-1}R'H_0R\left(R'R\right)^{-1}\bar{v}_2-\Delta_{n,\theta_0}'R\left(R'R\right)^{-1}\bar{v}_2+\bar{v}_2'\left(R'R\right)^{-1}RH_0h^+\leqslant o_P(1). \tag{48}$$

Using the definition of h^+ this reduces to

$$\frac{1}{2}\bar{v}_2'\left(R'R\right)^{-1}R'H_0R\left(R'R\right)^{-1}\bar{v}_2\leqslant o_P(1),$$

or $\bar{v}_2=o_P(1)$, so that $\bar{v}=o_P(1)$, $\bar{\eta}=o_P(1)$, $\bar{h}=h^+ +o_P(1)$.

Next we allow for singular H_0 in Theorem 9.1 of NM 1994. Note that by consistency, $g\left(\theta_0+h^*/\sqrt{n}\right)=0$ implies that $\bar{G}'h^*=0$, where $\bar{G}=G_0+o_P(1)$. We can then construct $\bar{R}$, continuously as a function of $\bar{G}$, such that $\bar{R}'\bar{G}=0$. This is possible since $\bar{R}$ can be the basis of the null space of $\bar{R}$, whose construction through the Gauss-Jordan process to an Echelon form is easily seen to be a continuous function. By the continuous mapping theorem, $\bar{R}=R+o_P(1)$, and since $\bar{B}=\left(\bar{G},\bar{R}\right)'$ is nonsingular w.p.c.1, $\bar{B}^{-1}=B^{-1}+o_P(1)$.

The same arguments as in Amemiya (1985) (pp21) for constructing h^+ can be applied to h^*:

$$h^* = \bar{R}\left(\bar{R}'H_0\bar{R}\right)^{-1}\bar{R}'\Delta_{n,\theta_0} = h^+ + o_P(1). \tag{49}$$

Finally, we resolve of the requirement of nonsingular H_0 in Theorem 1 of Sherman (1993). Replace (46) by (without using knowledge of $\bar{h} = O_P(1)$)

$$\Delta'_{n,\theta_0}\bar{h} - \frac{1}{2}\bar{h}'H_0\bar{h} \geqslant \Delta'_{n,\theta_0}h^+ - \frac{1}{2}h^{+'}H_0h^+ - o_P\left(1 + \|\bar{h}\|^2\right) \tag{50}$$

Now let $\bar{v} = \bar{B}\left(\bar{h} - h^+\right)$ for $\bar{B} = \left(\bar{G}, \bar{R}\right)'$, $\bar{G}'\bar{h} = 0$, $\bar{B} = B + o_P(1)$. The same manipulation above (replace (G_0, R) by $\left(\bar{G}, \bar{R}\right)$ if necessary) then shows that

$$\frac{1}{2}\bar{v}_2'\left(R'R\right)^{-1}R'H_0R\left(R'R\right)^{-1}\bar{v}_2 \leqslant o_P\left(1 + \|\bar{v}_2\|^2\right) \tag{51}$$

which also implies $\bar{v}_2 = o_P(1)$, and hence $h^* = h^+ + o_P(1) = O_P(1)$. Note $\bar{v}_1 = G_0'\bar{h}$ and $\bar{G}'\bar{h} = \left(G_0 + o_P(1)\right)'\bar{h} = 0$ imply that $G_0'\bar{h} = o_P\left(\bar{h}\right)$, or that $\|\bar{v}_1\|^2 = o_P\left(\|\bar{v}_1\|^2 + \|\bar{v}_2\|^2\right)$. Therefore $\|\bar{v}_1\| = o_P\left(\|\bar{v}_2\|\right)$. Finally, taylor expanding $g\left(\theta_0 + \bar{h}/\sqrt{n}\right) = 0$ with $\bar{h} = O_P(1)$ in turn implies $\bar{v}_1 = G_0'\bar{h} = O_p\left(\frac{1}{\sqrt{n}}\right)$. ∎

Proof of Theorem 3.

Define $u \equiv (u_1, u_2) = B\sqrt{n}\left(\theta - \bar{\theta}\right)$, representing directions orthogonal to and along the constrained subspace. Also let $\bar{h} = \sqrt{n}\left(\bar{\theta} - \theta_0\right)$. Then by Assumption 5 and Theorem 2, uniformly in $\|u/\sqrt{n}\| \leqslant o(1)$,

$$\begin{aligned} n\left(\hat{Q}_n\left(\bar{\theta} + B^{-1}\frac{u}{\sqrt{n}}\right) - \hat{Q}_n\left(\bar{\theta}\right)\right) &= -\frac{1}{2}u'\left(B^{-1}H_0B^{-1}\right)u \\ &\quad + \Delta'_{n,\theta_0}B^{-1}u - \bar{h}'H_0B^{-1}u + o_P\left(1 + \|u\|^2\right) \\ &= -\frac{1}{2}u'\left(B^{-1}H_0B^{-1}\right)u + \Delta'_{n,\theta_0}FG_0\left(G_0'G_0\right)^{-1}u_1 + o_P\left(1 + \|u\|^2\right) \end{aligned} \tag{52}$$

for $F = I - R\left(R'H_0R\right)^{-1}R'H_0$. This applies the local expansion from Assumption 5 to

$$\hat{Q}_n(\bar{\theta} + B^{-1}un^{-1/2}) = \hat{Q}_n(\theta_0 + \bar{h}n^{-1/2} + B^{-1}un^{-1/2}) \quad \text{and}$$
$$\hat{Q}_n(\bar{\theta}) = \hat{Q}_n(\theta_0 + \bar{h}n^{-1/2}).$$

Also, for $G^* = G_0 + O_P\left(u/\sqrt{n}\right)$, Taylor expand the penalties to write

$$\begin{aligned}
&\lambda_n^p \sqrt{n}^p \sum_{j=1}^{J} \left| g_j\left(\bar{\theta} + B^{-1}\frac{u}{\sqrt{n}}\right)\right|^p = \lambda_n^p \sum_{j=1}^{J} \left| G_j^{*\prime} B^{-1} u\right|^p \\
&= \lambda_n^p \sum_{j=1}^{J} \left| G_{0j}' B^{-1} u + O_p\left(u_j B^{-1} u/\sqrt{n}\right)\right|^p = \lambda_n^p \sum_{j=1}^{J} \left| u_{1j} + o_p\left(u_{1j}\right) + o_p\left(u_{2j}\right)\right|^p
\end{aligned} \tag{53}$$

Now, consider the case $p < \infty$. By the definition of $u^+ = B\sqrt{n}\left(\theta^+ - \bar{\theta}\right)$,

$$n\left(\bar{Q}_n\left(\bar{\theta} + B^{-1}u^+/\sqrt{n}\right) - \bar{Q}_n\left(\bar{\theta}\right)\right) \geqslant o_P\left(1\right).$$

where we have used $B^{-1} = \left(G_0\left(G_0'G_0\right)^{-1}\ R\left(R'R\right)^{-1}\right)$ and $G_{0j}'G_0\left(G_0'G_0\right)^{-1}u = u_{1j}$. Since $\theta^+ = \theta_0 + o_p(1)$ and $\bar{\theta} = \theta_0 + o_p(1)$ by Theorem 1, $\|u^+/\sqrt{n}\| = o_p(1)$. This together with (52) and (53) imply that w.p.c.1, $\exists \delta > 0$ such that

$$\delta\|u^+\|^2 - O_P\left(1\right)\|u_1^+\|_1 + \lambda_n^p\left\|u_1^+ + o_P\left(1\right)u_1^+ + o_P\left(1\right)u_2^+\right\|^p \leqslant o_P\left(1\right). \tag{54}$$

Given that the last term on the LHS is positive, (54) implies that

$$\delta\|u^+\|^2 - O_P\left(1\right)\|u_1^+\|_1 \leqslant o_P\left(1\right). \tag{55}$$

which in turn implies that $\|u^+\| = O_p\left(1\right)$. Then (54) implies $\|u_1^+\| = o_P\left(1\right)$. If not, then since $\lambda_n^p \to \infty$, the LHS is larger than any fixed number infinitely often with positive probability, contradicting (54). Finally, use (54) again to conclude that $\|u_2^+\| = o_P\left(1\right)$. Then $\sqrt{n}\left(\theta^+ - \bar{\theta}\right) = B^{-1}u^+ = o_P\left(1\right)$.

For the ℓ_∞ penalty, θ^+ satisfies

$$\delta\|u^+\|^2 - O_P(1)\|u_1^+\|_1 + \infty 1\left(\sum_{j=1}^{J}\left|u_{1j} + o_p(u_1) + o_p(u_2)\right| \geqslant 1\right) \leqslant o_P(1). \tag{56}$$

By (55) $\|u^+\| = o_P(1)$. Then it must be that $\|u_1^+\| = o_P(1)$. Otherwise the left hand side of (56) is ∞ infinitely often with positive probability and contradicts (56). ■

In Example 6, $\bar{\theta}_2 = \bar{X}$, $\bar{\theta}_1 = -\bar{X}^2$, For all $p < \infty$, $\theta^+ = \bar{\theta}$. For $p = \infty$, $\theta_2^+ = \bar{X}$, $\theta_1^+ \in \left(-\bar{X}^2 - \frac{1}{\lambda_n\sqrt{n}}, -\bar{X}^2 - \frac{1}{\lambda_n\sqrt{n}}\right)$, $\sqrt{n}\left(\theta_2^+ - \theta_{20}\right) \xrightarrow{d} N(0,1)$. For all $p < \infty$, $\sqrt{n}\left(\theta_1^+ - \theta_{10}\right) = o_P(1)$. For $p = \infty$, $\theta_1^+ - \theta_{10} = o_P(1)$ if $\lambda_n\sqrt{n} \to \infty$, and $\sqrt{n}\left(\theta_1^+ - \theta_{10}\right) = o_P(1)$ if $\lambda_n \to \infty$. Given data, the posterior distribution satisfies $\theta_2 \sim \pi(\theta_2) N\left(\bar{X}, \frac{1}{n}\right)$, $\theta_1 = u - \theta_2^2$ and $u \sim \pi(u) e^{-\lambda_n^p \sqrt{n}^p |u|^p}$. Posterior consistency is implied by $\lambda_n\sqrt{n} \to \infty$. For $p = \infty$ and uniform $\pi(u)$, $u \sim$ Uniform $\left(-\frac{1}{\lambda_n\sqrt{n}}, \frac{1}{\lambda_n\sqrt{n}}\right)$.

Proof of Theorem 4. Let $\hat{\pi}_0(v) = \pi_0\left(\bar{\theta} + B^{-1}D_n^{-1}v/\sqrt{n}\right)$, and $\pi_0 = \pi_0(\theta_0)$. Denote $H_n = D_n B\sqrt{n}\left(\Theta - \bar{\theta}\right)$,

$$w(v) = n\left(\hat{Q}_n\left(\bar{\theta} + B^{-1}D_n^{-1}v/\sqrt{n}\right) - \hat{Q}_n\left(\bar{\theta}\right)\right) + \sum_{j=1}^{J}\left|\lambda_n\sqrt{n}g_j\left(\bar{\theta} + B^{-1}D_n^{-1}\frac{v}{\sqrt{n}}\right)\right|^p \tag{57}$$

Then we can write, with $\bar{p}_v(v|\mathcal{X}_n) = \hat{\pi}_0(v)\exp(w(v))\,1(v \in H_n)$,

$$p_v(v|\mathcal{X}_n) = \frac{\bar{p}_v(v|\mathcal{X}_n)}{C_n} \quad \text{where} \quad C_n = \int_{H_n} \bar{p}_v(v|\mathcal{X}_n)\,dv. \tag{58}$$

We will show that for any finite $\alpha > 0$,

$$A_n = \int \|v\|^\alpha |\bar{p}_v(v|\mathcal{X}_n) - \bar{p}_v^\infty(v)|dv = o_P(1), \tag{59}$$

where for $v_1 \in \mathbb{R}^J$ and $v_2 \in \mathbb{R}^{K-J}$, $\bar{p}_v^\infty(v) = \pi_0 e^{-\frac{1}{2}v_2'\Sigma^{-1}v_2 - \sum_{j=1}^J |v_{1j}|^p}$. Also let $C_\infty = \int \bar{p}_v^\infty(v)\,dv = \pi_0(2\pi)^{\frac{K-J}{2}}\det|\Sigma|^{1/2}C_\kappa$.

Showing (59) is sufficient to prove the theorem. To see this, note that, for $p_v^\infty(v) =$

$\bar{p}_v^{\infty}(v)/C_{\infty}$,

$$\int \|v\|^{\alpha} |p_v(v|\mathcal{X}_n) - p_v^{\infty}(v)| dv = B_n C_n^{-1}, \tag{60}$$

where we can bound

$$B_n = \int \|v\|^{\alpha} \left| \bar{p}_v(v|\mathcal{X}_n) - \frac{C_n}{C_{\infty}} \bar{p}_v^{\infty}(v) \right| dv \leqslant A_n + \left| \frac{C_n - C_{\infty}}{C_{\infty}} \right| \int \|v\|^{\alpha} \bar{p}_v^{\infty}(v)\, dv.$$

The second term on the right-hand side is $o_P(1)$ because by (59), $|C_n - C_{\infty}| = o_P(1)$.

As a preliminary step to establishing (59), we show that it is enough to show convergence when the integral over v in the total variation of moments norm is restricted to a certain n-dependent subset of $\mathbb{R}^K$. Given any $\delta > 0$, find $\bar{\delta} > 0$, such that w.p.c.1,

$$\|D_n^{-1} v\| > \sqrt{n}\delta \Longrightarrow \|\theta - \bar{\theta}\| > 2\bar{\delta} \Longrightarrow \|\theta - \theta_0\| > \bar{\delta},$$

Then just as in the proof of Theorem 1, for any $\bar{\delta} > 0$, there exists $\eta(\bar{\delta}) > 0$, such that

$$\|\theta - \theta_0\| > \bar{\delta} \quad \text{implies} \quad Q^+(\theta) < Q^+(\theta_0) - \eta(\bar{\delta})$$

Note also that $\bar{\theta} = \theta_0 + O_P\left(\frac{1}{\sqrt{n}}\right)$, $g(\bar{\theta}) = 0$ imply that $|\bar{Q}_n^+(\theta^+) - \bar{Q}_n^+(\theta_0)|$ satisfies

$$b_n^{-1} |Q(\theta^+) - Q(\theta_0)| + o_P(1) = o_P\left(\max\left(\frac{1}{\sqrt{n}}, \frac{\sqrt{n}}{\lambda_n^p \sqrt{n^p}}\right)\right) + o_P(1) = o_P(1).$$

Then on an event sequence such that $\|\theta - \theta_0\| > \bar{\delta}$, w.p.c.1,

$$e^{n\left(\bar{Q}_n(\theta) - \bar{Q}_n(\bar{\theta})\right)} = e^{nb_n\left(\bar{Q}_n^+(\theta) - \bar{Q}_n^+(\theta_0) + o_P(1)\right)} \leqslant C_1 e^{-nb_n\eta(\bar{\delta})/2},$$

and hence,

$$\int_{\|D_n^{-1}v\| > \sqrt{n}\delta} \|v\|^{\alpha} \bar{p}_v(v|\mathcal{X}_n)\, dv \leqslant C e^{-nb_n\eta(\bar{\delta})/2} \int_{\|\theta - \theta_0\| > \bar{\delta}} \sqrt{n}^{K+\alpha} \lambda_n^{J+\alpha} \|\theta - \theta_0\|^{\alpha} \pi_0(\theta)\, d\theta = o_P(1).$$

Furthermore, it also holds that for any $M_n \to \infty$

$$\int_{\|v\| \geqslant M_n} \|v\|^{\alpha} \bar{p}_v^{\infty}(v)\, dv = o_P(1) \quad \text{so that} \quad \int_{\|D_n^{-1} v\| \geqslant \sqrt{n}\delta} \|v\|^{\alpha} \bar{p}_v^{\infty}(v)\, dv = o_P(1), \tag{61}$$

since $D_n^{-1} = (\lambda_n^{-1} I_J, I_{K-J})$ and $\bar{p}_v^{\infty}(v)$ has exponential tails.

First consider the case of $p < \infty$. Let $\bar{h} = \sqrt{n}\left(\bar{\theta} - \theta_0\right)$, $h = B^{-1} D_n^{-1} v$ and $F = I - R\left(R' H_0 R\right)^{-1} R' H_0$. By Assumption 5, for any $\delta \to 0$ sufficiently slowly, uniformly in v such that $\|D_n^{-1} v / \sqrt{n}\| < \delta$,

$$\begin{aligned} & n\left(\hat{Q}_n\left(\bar{\theta} + B^{-1} D_n^{-1} v/\sqrt{n}\right) - \hat{Q}_n\left(\bar{\theta}\right)\right) \\ =& \Delta'_{n,\theta_0} B^{-1} D_n^{-1} v - \bar{h}' H_0 B^{-1} D_n^{-1} v - \frac{1}{2} v' D_n^{-1} B^{-1} H_0 B^{-1} D_n^{-1} v + o_P\left(1 + \|h\|^2\right) \\ =& -\frac{1}{2} v' D_n^{-1} B^{-1} H_0 B^{-1} D_n^{-1} v + \Delta'_{n,\theta_0} F G_0 \left(G_0' G_0\right)^{-1} \frac{v_1}{\lambda_n} + o_P\left(1 + \|h\|^2\right), \end{aligned} \tag{62}$$

as in (52). Use Assumption 2 and $G_{0j}^* = G_{0j} + O_P\left(D_n^{-1} v/\sqrt{n}\right)$ to write

$$\begin{aligned} \lambda_n^p \sqrt{n}^p \sum_{j=1}^{J} \left|g_j\left(\bar{\theta} + B^{-1} D_n^{-1} v/\sqrt{n}\right)\right|^p &= \sum_{j=1}^{J} \left|\lambda_n G_{0j}^{*\prime} B^{-1} D_n^{-1} v\right|^p \\ = \sum_{j=1}^{J} \left| v_{1j} + o_P(1) + O_P\left(\frac{\lambda_n}{\sqrt{n}} \|v_2\|^2\right) + O_P\left(\frac{\|v_1\|^2}{\sqrt{n}\lambda_n}\right) \right|^p. \end{aligned} \tag{63}$$

This follows from $B^{-1} = \left(G_0\left(G_0' G_0\right)^{-1} \;\; R\left(R'R\right)^{-1}\right)$, $G_{0j}' R = 0$ and $G_{0j}' G_0 \left(G_0' G_0\right)^{-1} v = v_{1j}$. Because of (61) (with M_n the diameter of H_n), we can focus on showing

$$\begin{aligned} A_{n1} &= \int_{\|B^{-1} D_n^{-1} v\| \leqslant \sqrt{n}\delta} \|v\|^{\alpha} \left|\bar{p}_v(v|\mathcal{X}_n) - \bar{p}_v^{\infty}(v)\, 1(v \in H_n)\right| dv \\ &= \int_{\{v \in H_n, \|B^{-1} D_n^{-1} v\| \leqslant \sqrt{n}\delta\}} \|v\|^{\alpha} \bar{p}_v^{\infty}(v) \left| \frac{\pi_0\left(\bar{\theta} + B^{-1} D_n^{-1} v\right)}{\pi_0(\theta_0)} e^{\psi(v)} - 1 \right| dv = o_P(1). \end{aligned} \tag{64}$$

where we use (62) and (63) to write

$$\begin{aligned}\psi(v) =& \Delta'_{n,\theta_0} F G_0 (G_0'G_0)^{-1} \frac{v_1}{\lambda_n} - v_2' (R'R)^{-1} R' H_0 G_0 (G_0'G_0)^{-1} \frac{v_1}{\lambda_n} \\ & - \frac{1}{2\lambda_n^2} v_1' (G_0'G_0)^{-1} G_0' H_0 G_0 (G_0'G_0)^{-1} v_1 \\ & - \sum_{j=1}^{J} \left| v_{1j} + o_P(1) + O_P\left(\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2\right) + O_P\left(\frac{\|v_1\|^2}{\sqrt{n}\lambda_n}\right) \right|^p + \sum_{j=1}^{J} |v_{1j}|^p. \end{aligned} \tag{65}$$

For $M_n \to \infty$ sufficiently slowly,

$$\sup_{\|v\| \leqslant M_n} |\psi(v)| = o_P(1) \quad \text{and} \quad \sup_{\|v\| \leqslant M_n} \left| \frac{\pi_0\left(\bar{\theta} + B^{-1} D_n^{-1} v\right)}{\pi_0(\theta_0)} e^{\psi(v)} - 1 \right| = o_P(1). \tag{66}$$

Therefore,

$$A_{n11} = \int_{\{\|v\| \leqslant M_n, H_n, \|B^{-1} D_n^{-1} v\| \leqslant \delta\}} \|v\|^{\alpha} \bar{p}_v^{\infty}(v) \left| \frac{\pi_0\left(\bar{\theta} + B^{-1} D_n^{-1} v\right)}{\pi_0(\theta_0)} e^{\psi(v)} - 1 \right| dv = o_P(1). \tag{67}$$

Because of (61), to prove (64) it only remains to show that for any $M_n \to \infty$,

$$\int_{\|v\| \geqslant M_n, v \in H_n, \|B^{-1} D_n^{-1} v\| \leqslant \sqrt{n}\delta} \|v\|^{\alpha} \bar{p}_v(v|\mathcal{X}_n)\, dv = o_P(1). \tag{68}$$

Using (62) and (63), write $\bar{p}_v(v|\mathcal{X}_n) = \hat{\pi}_0(v) \exp(w(v))$ for

$$\begin{aligned} w(v) =& -\frac{1}{2} v' D_n^{-1} B^{-1} H_0 B^{-1} D_n^{-1} v + O_P\left(\frac{\|v_1\|}{\lambda_n}\right) + o_P(1) + o_P\left(\|v_2\|^2\right) + o_P\left(\frac{\|v_1\|^2}{\lambda_n^2}\right) \\ & - \sum_{j=1}^{J} \left| v_{1j} + o_P(1) + O_P\left(\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2\right) + O_P\left(\frac{\|v_1\|^2}{\sqrt{n}\lambda_n}\right) \right|^p. \end{aligned}$$

For some $\delta_k > 0$ denoting generic small constants, we can let $\|v_1\| \leqslant \delta_2 \lambda_n \sqrt{n}$ for any $\delta_2 > 0$ and n sufficiently large. There are two cases to consider. First, suppose on the previous event sequence that $\delta_3 \|v_1\| - \frac{\lambda_n}{\sqrt{n}} \|v_2\|^2 \to c \in [0, \infty]$. Then $w(v)$ is bounded above by

$$-\delta_1 \|v_2\|^2 - \sum_{j=1}^{J} \left| (1 - \delta_2 - \delta_3) v_{1j} \right|^p + o_P(1) \quad \text{w.p.c.1}$$

Second, suppose instead $\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2 - \delta_3\|v_1\| \to c \in (0,\infty]$. Then we replace the upper bound with

$$-\frac{\delta_1}{4}\|v_2\|^2 - \frac{\delta_1\delta_3}{4}\frac{\sqrt{n}}{\lambda_n}\|v_1\| + o_P(1) \quad \text{w.p.c.1.} \tag{69}$$

In either case, (68) holds because its left-hand side is $O_P\left(M_n^{\eta_1}e^{-\eta_2 M_n}\right) = o_P(1)$ for some $\eta_1, \eta_2 > 0$.

Finally, for the case $p = \infty$, replace (63) by $\infty 1\left(\left\|v_1 + O_P\left(\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2 + \frac{\|v_1\|^2}{\sqrt{n}\lambda_n}\right)\right\| \leqslant 1\right)$ and the remainder term in (65) by

$$\psi(v) = o_P\left(1 + \|v_2\|^2 + \frac{\|v_1\|^2}{\lambda_n^2}\right) - \infty 1\left(\left\|v_1 + O_P\left(\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2 + \frac{\|v_1\|^2}{\sqrt{n}\lambda_n}\right)\right\| \geqslant 1\right) + \infty 1\left(\|v_1\| \geqslant 1\right),$$

For any $\delta_n = o(1)$, it is clear that the integral in (64) over $|\|v_1\| - 1| \leqslant \delta_n$ is $o_P(1)$. It can then be shown that

$$\sup_{\|v\| \leqslant M_n, |\|v_1\| - 1| \geqslant \delta} |\psi(v)| = o_P(1)$$

since, for example

$$P\left(\|v_1\| \geqslant 1 + \delta, \left\|v_1 + O_P\left(\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2 + \frac{\|v_1\|^2}{\sqrt{n}\lambda_n}\right)\right\| < 1\right) = o(1).$$

Then (66) and (67) both hold. To show (68), if $\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2 \leqslant \delta_3\|v_1\|$, we bound w.p.c.1,

$$w(v) \leqslant -\delta_1\|v_2\|^2 - \infty 1\left(\|(1 - \delta_2 - \delta_3)v_1\| \geqslant 1\right) + o_P(1) \quad \text{w.p.c.1}$$

When $\frac{\lambda_n}{\sqrt{n}}\|v_2\|^2 > \delta_3\|v_1\|$, we then replace the upper bound by (69). The same $O_P\left(M_n^{\eta_1}e^{-\eta_2 M_n}\right)$ bound on (68) still holds. ∎

In Example 6, the localized posterior distribution $h = \sqrt{n}\left(\theta - \bar{\theta}\right)$ is proportional to

$$\exp\left(-\frac{1}{2}h_2^2 - \left|\lambda_n h_1 + 2\lambda_n \bar{X} h_2 + \frac{\lambda_n}{\sqrt{n}}h_2^2\right|^p\right)$$

For this to be approximated by $\exp\left(-\frac{1}{2}h_2^2 - \left|\lambda_n h_1\right|^p\right)$ it is necessary that $\frac{\lambda_n}{\sqrt{n}} \to 0$ so that

the posterior distribution can inform inference about $\sqrt{n}\left(\bar{\theta}-\theta_0\right)$.

PROOF OF THEOREM 5. We first show part (a). The result for the median (second part of (13)) will follow from (14) with $q_{1/2}=0$. For the posterior mean, define $\bar{\phi}(u)=\phi\left(\bar{\theta}+B^{-1}u\right)$, and write

$$\begin{aligned}\sqrt{n}\left(\phi^*-\phi\left(\bar{\theta}\right)\right) &= \int\sqrt{n}\left(\bar{\phi}\left(D_n^{-1}v/\sqrt{n}\right)-\bar{\phi}(0)\right)p_v\left(v|\mathcal{X}_n\right)dv \\ &= \frac{\partial}{\partial u_2'}\bar{\phi}(0)\int v_2 p_v\left(v|\mathcal{X}_n\right)dv+\frac{\partial}{\partial u_1'}\bar{\phi}(0)\int\frac{v_1}{\lambda_n}p_v\left(v|\mathcal{X}_n\right)dv+C_n.\end{aligned}$$

This is $o_P(1)$ because $\int v_2 p_v\left(v|\mathcal{X}_n\right)dv=o_P(1)$ and $\int v_1 p_v\left(v|\mathcal{X}_n\right)dv=o_P(1)$ by Theorem 4, and

$$|C_n|\leqslant C\int\left(\frac{\|v_1\|^2}{\lambda_n^2\sqrt{n}}+\frac{\|v_2\|^2}{\sqrt{n}}\right)p_v\left(v|\mathcal{X}_n\right)dv=o_p(1).$$

Next, we show the following conditional posterior weak convergence Delta method

$$\sqrt{n}\left(\phi(\theta)-\phi\left(\bar{\theta}\right)\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}N\left(0,\Lambda' R\left(R'HR\right)^{-1}R'\Lambda\right) \tag{70}$$

where $\theta\sim p_\theta\left(\theta|\mathcal{X}_n\right)$ and $\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}$ denotes conditional weak convergence in probability.

It follows from Theorem 4 that

$$\sqrt{n}\left(\theta-\bar{\theta}\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}B^{-1}\begin{pmatrix}0_J\\ N(0,\Sigma)\end{pmatrix}$$

It then follows from the conditional Delta method (e.g. Theorem 3.9.11 in van der Vaart and Wellner (1996), Lemma A.1 and Theorem 3.1 in Hong and Li (2018)) that

$$\sqrt{n}\left(\phi(\theta)-\phi\left(\bar{\theta}\right)\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}\Lambda' B^{-1}\begin{pmatrix}0_J\\ N(0,\Sigma)\end{pmatrix}$$

which is (70). That (70) implies (14) follows a probablistic version of Lemma 1.2.1 of Politis et al. (1999). Note that for $F_{n,\phi}(s)=P(\sqrt{n}(\phi(\theta)-\phi(\bar{\theta}))\leqslant s\mid\mathcal{X}_n)$, $\sqrt{n}\left(\phi_\tau^*-\phi\left(\bar{\theta}\right)\right)=F_{n,\phi}^{-1}(\tau)$

and $F_{\infty,\phi}^{-1}(\tau) = q_\tau \sqrt{\Lambda' R (R'HR)^{-1} R'\Lambda}$. Since $F_{\phi,\infty}(s)$ is strictly increasing in s, $\forall \epsilon > 0$, $\exists \delta > 0$ such that $F_{\phi,\infty}\left(F_{\phi,\infty}^{-1}(\tau) - \epsilon\right) \leqslant \tau - \delta$ and $F_{\phi,\infty}\left(F_{\phi,\infty}^{-1}(\tau) + \epsilon\right) \geqslant \tau + \delta$. Furthermore, $|F_{n,\phi}^{-1}(\tau) - F_{\phi,\infty}^{-1}(\tau)| > \epsilon$ implies either

$$F_{\phi,n}\left(F_{\phi,\infty}^{-1}(\tau) - \epsilon\right) \geqslant \tau \Longrightarrow F_{\phi,n}\left(F_{\phi,\infty}^{-1}(\tau) - \epsilon\right) - F_{\phi,\infty}\left(F_{\phi,\infty}^{-1}(\tau) - \epsilon\right) \geqslant \tau$$

or $F_{\phi,n}\left(F_{\phi,\infty}^{-1}(\tau) + \epsilon\right) \leqslant \tau \Longrightarrow F_{\phi,n}\left(F_{\phi,\infty}^{-1}(\tau) + \epsilon\right) - F_{\phi,\infty}\left(F_{\phi,\infty}^{-1}(\tau) + \epsilon\right) \geqslant \delta$. The probabilities of both events are $o(1)$. Thus (14) is proven.

Next we show (15). By Theorem 2 and the delta method,

$$\sqrt{n}\left(\phi\left(\bar{\theta}\right) - \phi(\theta_0)\right) = \Lambda' R (R'H_0R)^{-1} R'\Delta_{n,\theta_0} + o_P(1).$$

Then for each $\tau \in (0,1)$,

$$P\left(\phi_\tau^* \leqslant \phi(\theta_0)\right) = P\left(\sqrt{n}\left(\phi_\tau^* - \phi(\theta_0)\right) \leqslant 0\right) = P\left(\sqrt{n}\left(\phi_\tau^* - \phi\left(\bar{\theta}\right)\right) + \sqrt{n}\left(\phi\left(\bar{\theta}\right) - \phi(\theta_0)\right) \leqslant 0\right)$$
$$= P\left(\Lambda' R (R'H_0R)^{-1} R'\Delta_{n,\theta_0} + o_P(1) \leqslant -q_\tau \sqrt{\Lambda' R (R'H_0R)^{-1} R'\Lambda}\right) = (1-\tau) + o(1),$$

since when $R'\Omega R = R'H_0R$, $\Lambda' R (R'H_0R)^{-1} R'\Delta_{n,\theta_0} \rightsquigarrow N\left(0, \Lambda' R (R'H_0R)^{-1} R'\Lambda\right)$.

Next we consider the second part, where $\Lambda = G_0'\eta$ (so that $\Lambda' R = 0$). Note that $\frac{\partial}{\partial u_2'}\bar{\phi}(0) = \Lambda' R (R'R)^{-1} + O_P\left(\frac{1}{\sqrt{n}}\right) = O_P\left(\frac{1}{\sqrt{n}}\right)$, $\frac{\partial}{\partial u_1'}\bar{\phi}(0) = \eta' + O_P\left(\frac{1}{\sqrt{n}}\right)$. Therefore,

$$\begin{aligned} \lambda_n\sqrt{n}\left(\phi^* - \phi\left(\bar{\theta}\right)\right) &= \int \lambda_n\sqrt{n}\left(\bar{\phi}\left(D_n^{-1}v/\sqrt{n}\right) - \bar{\phi}(0)\right) p_v(v|\mathcal{X}_n)\, dv \\ &= \lambda_n \frac{\partial}{\partial u_2'}\bar{\phi}(0) \int v_2 p_v(v|\mathcal{X}_n)\, dv + \frac{\partial}{\partial u_1'}\bar{\phi}(0) \int v_1 p_v(v|\mathcal{X}_n)\, dv + C_n \end{aligned} \tag{71}$$

where the 1st term is $O_P\left(\frac{\lambda_n}{\sqrt{n}}\right) o_P(1) = o_P(1)$, 2nd term $(\eta' + o_P(1))\, o_P(1) = o_P(1)$, and

$$|C_n| \leqslant C \int \left(\frac{\|v_1\|^2}{\lambda_n\sqrt{n}} + \frac{\lambda_n}{\sqrt{n}}\|v_2\|^2\right) p_v(v|\mathcal{X}_n)\, dv = o_P(1)$$

This proves the claim for the posterior mean part of (16). The part about the posterior

quantiles will follow from

$$\lambda_n\sqrt{n}\left(\phi\left(\theta\right)-\phi\left(\bar{\theta}\right)\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}\eta'V_1. \tag{72}$$

which is essentially a multivariate conditional Delta method with differing convergence rates.

By the proof of Theorem 2, $G_0'\bar{h}=G_0'\sqrt{n}\left(\bar{\theta}-\theta_0\right)=O_p\left(\frac{1}{\sqrt{n}}\right)$. Hence $U_{1n}=\lambda_n\sqrt{n}G_0'\left(\bar{\theta}-\theta_0\right)=o_P\left(1\right)$, namely, $U_{1n}\rightsquigarrow 0$. Also, $U_{2n}=\sqrt{n}R'\left(\bar{\theta}-\theta_0\right)=O_P\left(1\right)$. By Theorem 4, $V_{1n}=\lambda_n\sqrt{n}G_0'\left(\theta-\bar{\theta}_0\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}V_1$, and $V_{2n}=\sqrt{n}R'\left(\theta-\bar{\theta}_0\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}V_2$. Taylor expanding to the 2nd order shows that $\lambda_n\sqrt{n}\left(\phi\left(\theta\right)-\phi\left(\bar{\theta}\right)\right)=o_P\left(1\right)$. Hence it suffices for (72) to show

$$\lambda_n\sqrt{n}\left(\phi\left(\theta\right)-\phi\left(\theta_0\right)\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}\eta'V_1. \tag{73}$$

Define $g\left(u,v\right)=\eta'\left(u_1+v_1\right)$, and

$$g_n\left(u,v\right)=\lambda_n\sqrt{n}\left(\phi\left(\theta_0+B^{-1}\left(\frac{u_1'}{\lambda_n\sqrt{n}},\frac{u_2'}{\sqrt{n}}\right)'+B^{-1}\left(\frac{v_1'}{\lambda_n\sqrt{n}},\frac{v_2'}{\sqrt{n}}\right)'\right)-\phi\left(\theta_0\right)\right)$$

Then by a second order Taylor expansion, $g_n\left(u,v\right)\to g\left(u,v\right)$. Invoke the extended continuous mapping theorem (Theorem 1.11.1 van der Vaart and Wellner (1996) and Lemma A.1 Hong and Li (2018)) to claim (73):

$$g_n\left(U_{1n},U_{2n},V_{1n},V_{2n}\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}g\left(U_1,U_2,V_1,V_2\right)=\eta'\left(U_1+V_1\right)=\eta'V_1.$$

Note that for any $\tau\in\left(0,0.5\right)$, $\bar{q}_\tau<0$ by symmetry of $\eta'V_1$. Then for $a_n=\lambda_n\sqrt{n}$,

$$P\left(\phi_\tau^*\leqslant\phi\left(\theta_0\right)\right)=P\left(a_n\left(\phi_\tau^*-\phi\left(\bar{\theta}\right)\right)\leqslant a_n\left(\phi\left(\theta_0\right)-\phi\left(\bar{\theta}\right)\right)\right)=P\left(\bar{q}_\tau\leqslant o_P\left(1\right)\right)=1-o\left(1\right).$$

Similarly, for any $\tau>0.5$, $P\left(\phi_\tau^*\leqslant\phi\left(\theta_0\right)\right)=o\left(1\right)$. The proof is thus completed.

Finally, we also show (18). Note that

$$
\begin{aligned}
&nVar\left(\phi\left(\theta\right)|\mathcal{X}_n\right) = \int n\left(\phi\left(\theta\right)-\phi\left(\theta_0\right)\right)\left(\phi\left(\theta\right)-\phi\left(\theta_0\right)\right)' p\left(\theta|\mathcal{X}_n\right) d\theta + o_P\left(1\right) \\
&= \int n\left(\bar{\phi}\left(D_n^{-1}v/\sqrt{n}\right)-\bar{\phi}\left(0\right)\right)\left(\bar{\phi}\left(D_n^{-1}v/\sqrt{n}\right)-\bar{\phi}\left(0\right)\right)' p_v\left(v|\mathcal{X}_n\right) dv + o_P\left(1\right) \\
&= \frac{\partial\bar{\phi}\left(0\right)'}{\partial u_2}\int v_2 v_2' p_v\left(v|\mathcal{X}_n\right) dv \frac{\partial\bar{\phi}\left(0\right)}{\partial u_2} + \frac{\partial\bar{\phi}\left(0\right)'}{\partial u_1}\int \frac{v_1 v_1'}{\lambda_n^2} p_v\left(v|\mathcal{X}_n\right) dv \frac{\partial\bar{\phi}\left(0\right)}{\partial u_1} + C_n + o_P\left(1\right)
\end{aligned}
$$

where $\frac{\partial\bar{\phi}(0)'}{\partial u_2} = \Lambda' R\left(R'R\right)^{-1} + o_P\left(1\right)$, $\int v_2 v_2' p_v\left(v|\mathcal{X}_n\right) dv = \Sigma + o_P\left(1\right)$, and

$$
|C_n| \leqslant C\int\left(\frac{\|v_1\|^2}{\lambda_n^2\sqrt{n}} + \frac{\|v_2\|^2}{\sqrt{n}}\right)^2 p_v\left(v|\mathcal{X}_n\right) dv = o_P\left(1\right).
$$

Therefore (18) holds. ∎

Proof of Theorem 6. We first show that $\theta^+ = \theta_0 + o_p(1)$ by modifying the proof of Theorem 1(b). Recall the definition of $\epsilon(\delta)$ from that proof. Redefine $d_n = \bar{\lambda}_n^p\sqrt{n}^p/n = \lambda_n^p n^{\gamma p/2}/n$ and $b_n = \min(1, d_n)$. Specifically, we established two preliminary results in Theorem 1(b) that we will show to hold in the adaptive case. The first result holds in our new context with the following minor modification to (44):

$$
\sup_{\theta\in\Theta}|\bar{Q}_n^+\left(\theta\right) - Q^+\left(\theta\right)| = o_P\left(1\right) \quad \text{for} \quad Q^+\left(\theta\right) = b_n^{-1}Q\left(\theta\right) - b_n^{-1}\frac{\lambda_n^p}{n}\sum_{m=1}^{M}|\hat{w}_m g_m\left(\theta\right)|^p. \tag{74}
$$

For the second preliminary result, for any $\delta > 0$, we shall argue for the existence of $\eta\left(\delta\right) > 0$ and a positive sequence of possibly data-dependent terms $\{R_n : n \in \mathbb{N}\}$, such that

$$
\|\theta^+ - \theta_0\| > \delta \quad \text{implies} \quad Q^+\left(\theta^+\right) < Q^+\left(\theta_0\right) - \eta\left(\delta\right) R_n. \tag{75}
$$

On the event $\|\theta^+ - \theta_0\| > \delta$, either (1) $\|\theta_g\left(\theta^+\right) - \theta_0\| > \left(1 - \frac{1}{K}\right)\delta$ and $\|\theta_g\left(\theta^+\right) - \theta^+\| < \frac{1}{K}\delta$, or (2) $\|\theta_g\left(\theta^+\right) - \theta^+\| > \frac{1}{K}\delta$. In the first case, we can take, as in the proof of Theorem 1(b),

$$
\eta(\delta) = \epsilon((1 - K^{-1})\delta) - \epsilon(\delta K^{-1})
$$

and $R_n = 1$. In the second case (2), $|g_j\left(\theta^+\right)| \geqslant \epsilon\left(\frac{1}{K}\delta\right) \equiv \bar{\epsilon}\left(\delta\right)$ for all $j = 1, \ldots, J$. Since

$Q(\theta_0) \geqslant Q(\theta^+)$ by Assumption 4,

$$Q^+\left(\theta^+\right) - Q^+\left(\theta_0\right) \leqslant -b_n^{-1}\frac{\lambda_n^p}{n}\bar{\epsilon}\left(\delta\right)^p \sum_{j=1}^{J}|\tilde{g}_j|^{-\gamma p} \leqslant -\bar{\epsilon}\left(\delta\right)^p \sum_{j=1}^{J}|\sqrt{n}\tilde{g}_j|^{-\gamma p},$$

We then take $\eta(\delta) = \bar{\epsilon}(\delta)^p$ and $R_n = \sum_{j=1}^{J}|\sqrt{n}\tilde{g}_j|^{-\gamma p}$.

Finally, we prove the assertion using these preliminary results:

$$\begin{aligned} P\left(\|\theta^+ - \theta_0\| > \delta\right) &\leqslant P\left(Q^+\left(\theta^+\right) < Q^+\left(\theta_0\right) - \eta\left(\delta\right) R_n\right) \\ &= P\left(Q^+\left(\theta^+\right) - \bar{Q}_n^+\left(\theta^+\right) < Q^+\left(\theta_0\right) - \bar{Q}_n^+(\theta_0) + \bar{Q}_n^+(\theta_0) - \bar{Q}_n^+\left(\theta^+\right) - \eta\left(\delta\right) R_n\right) \\ &\leqslant P\left(\sup_{\theta\in\Theta}|\bar{Q}_n^+\left(\theta\right) - Q^+\left(\theta\right)| > \frac{\eta\left(\delta\right)}{3}R_n\right) + P\left(\bar{Q}_n^+\left(\theta^+\right) - \bar{Q}_n^+\left(\theta_0\right) < -\frac{\eta\left(\delta\right)}{3}R_n\right), \end{aligned}$$

where the first line follows from (75) and the last line from the law of total probability. It remains to show that the two terms on the last line are $o(1)$. In case (1) above where $R_n = 1$, this follows from the same arguments in the proof of Theorem 1(b). For case (2), note that by the law of total probability,

$$\begin{aligned} P\left(\sup_{\theta\in\Theta}|\bar{Q}_n^+\left(\theta\right) - Q^+\left(\theta\right)| > \frac{\eta\left(\delta\right)}{3}R_n\right) \leqslant P\left(\sup_{\theta\in\Theta}|\bar{Q}_n^+\left(\theta\right) - Q^+\left(\theta\right)| > \frac{\eta\left(\delta\right)}{3}JM^{-\gamma p}\right) \\ + J\max_{j\in\{1,\ldots,J\}} P\left(|\sqrt{n}\tilde{g}_j| > M\right). \end{aligned}$$

For any fixed M, the first term on the right-hand side is $o(1)$ by (74). On the other hand, the second term satisfies

$$\limsup_{M\to\infty}\lim_{n\to\infty} J\max_{j\in\{1,\ldots,J\}} P\left(|\sqrt{n}\tilde{g}_j| > M\right) = 0$$

by Assumption 8, noting that $g_j(\theta_0) = 0$ for all $j = 1, \ldots, J$. Likewise,

$$\begin{aligned} P\left(\bar{Q}_n^+\left(\theta^+\right) - \bar{Q}_n^+\left(\theta_0\right) < -\frac{\eta\left(\delta\right)}{3}R_n\right) \leqslant P\left(\bar{Q}_n^+\left(\theta^+\right) - \bar{Q}_n^+\left(\theta_0\right) < -\frac{\eta\left(\delta\right)}{3}JM^{-\gamma p}\right) \\ + J\max_{j\in\{1,\ldots,J\}} P\left(|\sqrt{n}\tilde{g}_j| > M\right), \end{aligned}$$

and the first term on the right-hand side is $o(1)$, since $\bar{Q}_n^+(\theta^+) \geqslant \bar{Q}_n^+(\theta_0) - o_P(1)$ by definition.

Now we show $\theta^+ - \bar{\theta} = o_p(n^{-1/2})$ by modifying the proof of Theorem 3. For the misspecified constraints $(g_m(\theta), m = J+1, \ldots, M)$, w.p.c.1, for $G^* = G_0 + O_P(u/\sqrt{n})$,

$$
\begin{aligned}
&\lambda_n^p \sum_{m=J+1}^{M} \frac{1}{|\tilde{g}_m|^{p\gamma}} \left(\left| g_m\left(\bar{\theta} + B^{-1}\frac{u}{\sqrt{n}}\right)\right|^p - |g_m(\bar{\theta})|^p \right) \\
&= \lambda_n^p \sum_{m=J+1}^{M} \left\{ \frac{1}{|\tilde{g}_m|^{p\gamma}} \left| G_m^{*\prime} B^{-1} u/\sqrt{n}\right|^p + \frac{C_1 |g_m(\bar{\theta})|}{|\tilde{g}_m|^{p\gamma}} \left| G_m^{*\prime} B^{-1} u/\sqrt{n}\right|^{p-1} \right. \\
&\qquad \left. + \frac{C_2 |g_m(\bar{\theta})|^2}{|\tilde{g}_m|^{p\gamma}} \left| G_m^{*\prime} B^{-1} u/\sqrt{n}\right|^{p-2} + \ldots + \frac{C_{p-1} |g_m(\bar{\theta})|^{p-1}}{|\tilde{g}_m|^{p\gamma}} \left| G_m^{*\prime} B^{-1} u/\sqrt{n}\right| \right\} \\
&= \lambda_n^p O_p \left(\sum_{m=J+1}^{M} \left| G_{0m}' B^{-1} u/\sqrt{n} + u' B^{-1} u/n \right| \right) \\
&= \frac{\lambda_n^p}{\sqrt{n}} O_p \left(\sum_{m=J+1}^{M} \left| u_{1m} + o_p\left(\|u\|^2\right)\right| \right) = o_p(1)
\end{aligned}
$$

by Assumption 9(d).

Also, (53) continues to hold for the correctly specified constraints $g_j(\theta), j = 1 \ldots J$:

$$
\begin{aligned}
&\lambda_n^p \sum_{j=1}^{J} \frac{1}{|\tilde{g}_j|^{p\gamma}} \left(\left| g_j\left(\bar{\theta} + B^{-1}\frac{u}{\sqrt{n}}\right)\right|^p \right) \\
&= \bar{\lambda}_n^p \sqrt{n}^p \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} \left(\left| g_j\left(\bar{\theta} + B^{-1}\frac{u}{\sqrt{n}}\right)\right|^p \right) \\
&= \bar{\lambda}_n^p \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} \left(\left| G_j^{*\prime} B^{-1} u\right|^p \right) \\
&= \bar{\lambda}_n^p \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} \left| u_{1j} + o_p\left(\|u\|^2\right)\right|^p
\end{aligned}
$$

Therefore, (54) can be replaced by

$$
\delta \|u^+\|^2 - O_P(1)\|u_1^+\|_1 + \bar{\lambda}_n^p \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} \left\| (1 + o_P(1)) u_1^+ + o_P(1) u_2^+ \right\|_1^p \leqslant o_P(1). \tag{76}
$$

Given that the last term on the LHS is positive,

$$\delta\|u^+\|^2 - O_P(1)\|u_1^+\|_1 \leqslant o_P(1).$$

The rest of proof of Theorem 3 then goes through.

■

PROOF OF THEOREM 7. Continuing from the proof of Theorem 4, replace (57) by

$$\begin{aligned} w(v) = n&\left(\hat{Q}_n\left(\bar{\theta} + \frac{B^{-1}\bar{D}_n^{-1}v}{\sqrt{n}}\right) - \hat{Q}_n\left(\bar{\theta}\right)\right) \\ &+ \log\kappa_p\left(\lambda_n\hat{w}\circ g\left(\bar{\theta} + B^{-1}\bar{D}_n^{-1}\frac{v}{\sqrt{n}}\right)\right) - \log\kappa_p\left(\lambda_n\hat{w}\circ g\left(\bar{\theta}\right)\right) \end{aligned}$$

We will show (59) with $p_v^\infty(v)$ replaced by

$$\bar{p}_v^\infty(v|\mathcal{X}_n) = \pi_0 e^{-\frac{1}{2}v_2'\Sigma^{-1}v_2 - \sum_{j=1}^J|\sqrt{n}\tilde{g}_j|^{-p\gamma}|v_{1j}|^p}$$

Also let $C_\infty = \int \bar{p}_v^\infty(v|\mathcal{X}_n)\,dv = \pi_0(2\pi)^{\frac{K-J}{2}}\det(\Sigma)^{1/2}\bar{C}_{\kappa_J}$. By Assumption 10, $C_\infty^{-1} = O_P(1)$, so that $C_n^{-1} = (C_\infty + o_P(1))^{-1} = O_P(1)$. Thus $B_n = o_P(1)$ in (60) follows from (59).

Given any $\delta > 0$, find $\bar{\delta} > 0$, such that w.p.c.1,

$$\|\bar{D}_n^{-1}v\| > \sqrt{n}\delta \Longrightarrow \|\theta - \bar{\theta}\| > 2\bar{\delta} \Longrightarrow \|\theta - \theta_0\| > \bar{\delta},$$

Then argue as in the proof of Theorem 6 that for any $\bar{\delta} > 0$, there exists $\eta(\bar{\delta}) > 0$ and a positive sequence of possibly data-dependent terms $\{R_n : n \in \mathbb{N}\}$, such that

$$\|\theta - \theta_0\| > \bar{\delta} \quad \text{implies} \quad Q^+(\theta) < Q^+(\theta_0) - \eta(\bar{\delta})R_n. \tag{77}$$

Note also that $\bar{\theta} = \theta_0 + O_P\left(\frac{1}{\sqrt{n}}\right)$, $\frac{\lambda_n^p}{n}\sum_{m=1}^M|\hat{w}_m g_m(\bar{\theta})|^p = o_P\left(\frac{1}{\sqrt{n}}\right)$ imply that $|\bar{Q}_n^+(\theta^+) - \bar{Q}_n^+(\theta_0)|$ satisfies

$$b_n^{-1}|Q(\theta^+) - Q(\theta_0)| + o_P(1) = o_P\left(\max\left(\frac{1}{\sqrt{n}}, \frac{\sqrt{n}}{\bar{\lambda}_n^p\sqrt{n^p}}\right)\right) + o_P(1) = o_P(1).$$

We can then write, for $\|\theta-\theta_0\|>\bar{\delta}$, w.p.c.1,

$$e^{n\left(\bar{Q}_n(\theta)-\bar{Q}_n\left(\bar{\theta}\right)\right)}=e^{nb_n\left(\bar{Q}_n^+(\theta)-\bar{Q}_n^+(\theta_0)+o_P(1)\right)}\leqslant C_1e^{-nb_n\eta\left(\bar{\delta}\right)R_n/2}.$$

This yields

$$\int_{\|\bar{D}_n^{-1}v\|>\sqrt{n}\delta}\|v\|^\alpha\bar{p}_v\left(v|\mathcal{X}_n\right)dv\leqslant Ce^{-nb_n\eta\left(\bar{\delta}\right)R_n/2}\int_{\|\theta-\theta_0\|>\bar{\delta}}\sqrt{n}^{K+\alpha}\bar{\lambda}_n^{J+\alpha}\|\theta-\theta_0\|^\alpha\pi_0\left(\theta\right)d\theta=o_P\left(1\right).$$

Furthermore, since $\bar{D}_n^{-1}=\left(\bar{\lambda}_n^{-1}I_J,I_{K-J}\right)$ and $\bar{p}_\infty\left(v|\mathcal{X}_n\right)$ has exponential tails, it also holds that for any $M_n\to\infty$,

$$\int_{\|v\|\geqslant M_n}\|v\|^\alpha\bar{p}_v^\infty\left(v|\mathcal{X}_n\right)dv=o_P\left(1\right)\quad\text{so that}\quad\int_{\|\bar{D}_n^{-1}v\|\geqslant\sqrt{n}\delta}\|v\|^\alpha\bar{p}_v^\infty\left(v|\mathcal{X}_n\right)dv=o_P\left(1\right).\tag{78}$$

By Assumption 5, for any $\delta\to0$ sufficiently slowly, uniformly in v such that $\|B^{-1}\bar{D}_n^{-1}v/\sqrt{n}\|\leqslant\delta$,

$$\begin{aligned}&n\left(\hat{Q}_n\left(\bar{\theta}+B^{-1}\bar{D}_n^{-1}v/\sqrt{n}\right)-\hat{Q}_n\left(\bar{\theta}\right)\right)\\&=-\frac{1}{2}v'\bar{D}_n^{-1}B^{-1}H_0B^{-1}\bar{D}_n^{-1}v+\Delta'_{n,\theta_0}FG_0\left(G_0'G_0\right)^{-1}\frac{v_1}{\bar{\lambda}_n}+o_P\left(1+\|h\|^2\right),\end{aligned}\tag{79}$$

where $h=B^{-1}\bar{D}_n^{-1}v$ and $F=I-R\left(R'H_0R\right)^{-1}R'H_0$.

Since $g_j\left(\bar{\theta}\right)=0$ for $j=1,\ldots,J$ and $G^*=G_0+O_P\left(\frac{B^{-1}\bar{D}_n^{-1}v}{\sqrt{n}}\right)$, the correctly specified

constraints satisfy

$$\begin{aligned}
&\lambda_n^p \sum_{j=1}^{J} \frac{1}{|\tilde{g}_j|^{p\gamma}} \left(\left| g_j \left(\bar{\theta} + B^{-1} \bar{D}_n^{-1} \frac{v}{\sqrt{n}} \right) \right|^p \right) \\
&= \bar{\lambda}_n^p \sqrt{n}^p \sum_{j=1}^{J} \left| \sqrt{n} \tilde{g}_j \right|^{-p\gamma} \left(\left| g_j \left(\bar{\theta} + B^{-1} \bar{D}_n^{-1} \frac{v}{\sqrt{n}} \right) \right|^p \right) \\
&= \sum_{j=1}^{J} \left| \sqrt{n} \tilde{g}_j \right|^{-p\gamma} \left(\left| \bar{\lambda}_n G_j^{*\prime} B^{-1} \bar{D}_n^{-1} v \right|^p \right) \\
&= \sum_{j=1}^{J} \left| \sqrt{n} \tilde{g}_j \right|^{-p\gamma} \left(\left| \bar{\lambda}_n G_{0j}^{\prime} B^{-1} \bar{D}_n^{-1} v + \bar{\lambda}_n v^{\prime} \bar{D}_n^{-1} B^{-1\prime} B^{-1} \bar{D}_n^{-1} v / \sqrt{n} \right|^p \right) \\
&= \sum_{j=1}^{J} \left| \sqrt{n} \tilde{g}_j \right|^{-p\gamma} \left| v_{1j} + O_p \left(\frac{\bar{\lambda}_n}{\sqrt{n}} \left\| v_2 \right\|^2 \right) + O_p \left(\frac{\left\| v_1 \right\|^2}{\sqrt{n} \bar{\lambda}_n} \right) \right|^p
\end{aligned}$$

In contrast, for the misspecified constraints,

$$\begin{aligned}
&\lambda_n^p \sum_{m=J+1}^{M} \frac{1}{|\tilde{g}_m|^{p\gamma}} \left(\left| g_m \left(\bar{\theta} + B^{-1} \bar{D}_n^{-1} \frac{v}{\sqrt{n}} \right) \right|^p - \left| g_m \left(\bar{\theta} \right) \right|^p \right) \\
&= \lambda_n^p \sum_{m=J+1}^{M} \left\{ \frac{1}{|\tilde{g}_m|^{p\gamma}} \left| G_m^* B^{-1} \bar{D}_n^{-1} v / \sqrt{n} \right|^p + \frac{C_1 \left| g_m \left(\bar{\theta} \right) \right|}{|\tilde{g}_m|^{p\gamma}} \left| G_m^* B^{-1} \bar{D}_n^{-1} v / \sqrt{n} \right|^{p-1} \right. \\
&\qquad \left. + \frac{C_2 \left| g_m \left(\bar{\theta} \right) \right|^2}{|\tilde{g}_m|^{p\gamma}} \left| G_m^* B^{-1} \bar{D}_n^{-1} v / \sqrt{n} \right|^{p-2} + ... + \frac{C_{p-1} \left| g_m \left(\bar{\theta} \right) \right|^{p-1}}{|\tilde{g}_m|^{p\gamma}} \left| G_m^* B^{-1} \bar{D}_n^{-1} v / \sqrt{n} \right| \right\} \\
&= \lambda_n^p O_p \left(\sum_{m=J+1}^{M} \left| G_{0m}^{\prime} B^{-1} \bar{D}_n^{-1} v / \sqrt{n} + v^{\prime} \bar{D}_n^{-1} B^{-1\prime} B^{-1} \bar{D}_n^{-1} v / n \right| \right) \\
&= \frac{\lambda_n^p}{\sqrt{n}} O_p \left(\sum_{m=J+1}^{M} \left| v_{1m} + O_p \left(\frac{\bar{\lambda}_n}{\sqrt{n}} \left\| v_2 \right\|^2 \right) + O_p \left(\frac{\left\| v_1 \right\|^2}{\sqrt{n} \bar{\lambda}_n} \right) \right| \right) = o_p(1)
\end{aligned}$$

by Assumption 9.

Using the previous two relations and (62), (64) holds with $\bar{p}_v^{\infty}(v)$ replaced by $\bar{p}_{\infty}(v|\mathcal{X}_n)$,

H_n replaced by $\bar{H}_n \equiv \bar{D}_n B\sqrt{n}\left(\Theta - \bar{\theta}\right)$, and

$$\begin{aligned}\psi(v) =& \Delta'_{n,\theta_0} F G_0 \left(G_0'G_0\right)^{-1} \frac{v_1}{\bar{\lambda}_n} - v_2' \left(R'R\right)^{-1} R' H_0 G_0 \left(G_0'G_0\right)^{-1} \frac{v_1}{\bar{\lambda}_n} \\ &- \frac{1}{2\bar{\lambda}_n^2} v_1' \left(G_0'G_0\right)^{-1} G_0' H F_0 G_0 \left(G_0'G_0\right)^{-1} v_1 \\ &- \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} \left| v_{1j} + O_p\left(\frac{\bar{\lambda}_n}{\sqrt{n}} \|v_2\|^2\right) + O_p\left(\frac{\|v_1\|^2}{\sqrt{n}\bar{\lambda}_n}\right)\right|^p + \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} |v_{1j}|^p ,\end{aligned}$$

which replaces (65). Furthermore, (66) and (67) continue to hold. Finally (68) also holds since we can now write

$$\begin{aligned}w(v) =& -\frac{1}{2} v' \bar{D}_n^{-1} B^{-1} H_0 B^{-1} \bar{D}_n^{-1} v - \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} \left| v_{1j} + O_p\left(\frac{\bar{\lambda}_n}{\sqrt{n}} \|v_2\|^2\right) + O_p\left(\frac{\|v_1\|^2}{\sqrt{n}\bar{\lambda}_n}\right)\right|^p \\ &+ o_P\left(\frac{\|v_1\|^2}{\bar{\lambda}_n^2}\right) + o_P\left(\|v_2\|^2\right) + o_P\left(\frac{\|v_1\|}{\bar{\lambda}_n}\right) + o_P(1)\end{aligned}$$

For some $\delta_k > 0$ denoting generic small constants, we can let $\|v_1\| \leqslant \delta_2 \bar{\lambda}_n \sqrt{n}$ for any $\delta_2 > 0$ and n sufficiently large. There are two cases to consider. First, suppose on the previous event sequence that $\delta_3 \|v_1\| - \frac{\lambda_n}{\sqrt{n}} \|v_2\|^2 \to \alpha \in [0, \infty]$. Then $w(v)$ is bounded above by

$$-\delta_1 \|v_2\|^2 - \sum_{j=1}^{J} \left|\sqrt{n}\tilde{g}_j\right|^{-p\gamma} \left| (1 - \delta_2 - \delta_3) v_{1j}\right|^p + o_P(1) \quad \text{w.p.c.1}$$

Second, suppose instead $\frac{\lambda_n}{\sqrt{n}} \|v_2\|^2 - \delta_3 \|v_1\| \to \alpha \in (0, \infty]$. Then we replace the upper bound with

$$-\frac{\delta_1}{4} \|v_2\|^2 - \frac{\delta_1 \delta_3}{4} \frac{\sqrt{n}}{\bar{\lambda}_n} \|v_1\| + o_P(1) \quad \text{w.p.c.1} \tag{80}$$

In either case, the left-hand side of (68) is $O_P\left(M_n^{\eta_1} e^{-\eta_2 M_n}\right) = o_P(1)$ for some $\eta_1, \eta_2 > 0$.

Finally, the proof for Theorem 5 goes through verbatim upon replacing λ_n with $\bar{\lambda}_n$ and $p_v^{\infty}(v)$ with $p_{\infty}(v|\mathcal{X}_n)$. ■

PROOF OF THEOREM 8. We first show consistency, $\bar{\theta}_S = \theta_0 + o_P(1)$, using arguments similar to the proof of Theorem 1, and we employ the notation defined there. As in that proof,

the event $\|\bar{\theta}_S - \theta_0\| < \delta$ can be bounded by the union of two events: (1) $\|\theta_g\left(\bar{\theta}_S\right) - \bar{\theta}_S\| \geqslant \delta/K$; and (2) $Q\left(\bar{\theta}_S\right) \leqslant Q\left(\theta_0\right) - \eta$. Event (1) has vanishing probability since $g\left(\cdot\right)$ is continuous and $g\left(\bar{\theta}_S\right) = o_P\left(1\right)$ by Assumption 11(a). Event (2) will also have vanishing probability if we can show $Q\left(\bar{\theta}_S\right) \geqslant Q\left(\theta_0\right) - o_P\left(1\right)$. For this purpose, let $\Theta_S = \{\theta \in \Theta : \|g_n\left(\theta\right)\| \leqslant \epsilon_n\}$ and note that by definition of $\bar{\theta}_S$ and Assumption 3, $Q\left(\bar{\theta}_S\right) \geqslant \sup_{\theta\in\Theta_S} Q\left(\theta\right) - o_P\left(1\right)$. It then suffices to show that $\sup_{\theta\in\Theta_S} Q\left(\theta\right) \geqslant Q\left(\theta_0\right) - o_P\left(1\right)$, which in turn follows from the continuity of $Q\left(\cdot\right)$ if we can show that $\inf_{\theta\in\Theta_S} \|\theta - \theta_0\| = o_P\left(1\right)$.

To show this, note that the constraint set Θ_S can be equivalently expressed as $\Theta_S = \left\{\theta \in \Theta : \|g_n\left(\theta\right)\|^2 \leqslant \epsilon_n^2\right\}$. Furthermore, if $\inf_{\|g_n(\theta)\|^2\leqslant\epsilon_n^2} \|\theta - \theta_0\|^2 = o_p(1)$, then $\inf_{\|g_n(\theta)\|^2\leqslant\epsilon_n^2} \|\theta - \theta_0\| = o_p(1)$ as well. Since the constraint set is convex and the objective function is convex, there exists a unique minimizer $\theta^* = \underset{\|g_n(\theta)\|^2\leqslant\epsilon_n^2}{\arg\min} \|\theta - \theta_0\|^2$. The Lagrangian is

$$L\left(\theta\right) = \|\theta - \theta_0\|^2 + \lambda\left(\|g_n\left(\theta\right)\|^2 - \epsilon_n^2\right).$$

The first order KKT conditions are, for $G_n\left(\theta\right) = \frac{\partial g_n(\theta)}{\partial \theta'}$,

$$\begin{aligned}
&\nabla L\left(\theta^*\right) = 2\left(\theta^* - \theta_0\right) + 2\lambda G_n\left(\theta^*\right) g_n\left(\theta^*\right) = 0, \\
&\lambda\left(\|g_n\left(\theta^*\right)\|^2 - \epsilon_n^2\right) = 0, \\
&\lambda \geqslant 0.
\end{aligned}$$

Taylor expanding the first KKT condition and using Assumption 11(b),

$$\begin{aligned}
&\left(\theta^* - \theta_0\right) + \lambda\left(G_n\left(\theta_0\right) + O_p\left(1/\sqrt{n}\right)\right)\left(g_n\left(\theta_0\right) + G_n\left(\theta_0\right)'\left(\theta^* - \theta_0\right) + O_p\left(1/\sqrt{n}\right)\right) = 0 \\
\Longrightarrow\ &\left(I + \lambda G_n\left(\theta_0\right) G_n\left(\theta_0\right)'\right)\left(\theta^* - \theta_0\right) = -\lambda G_n\left(\theta_0\right) g_n\left(\theta_0\right) + O_p\left(1/\sqrt{n}\right) \\
\Longrightarrow\ &\theta^* - \theta_0 = -\lambda\left(I + \lambda G_n\left(\theta_0\right) G_n\left(\theta_0\right)'\right)^{-1} G_n\left(\theta_0\right) g_n\left(\theta_0\right) + O_p\left(1/\sqrt{n}\right) = O_p\left(1/\sqrt{n}\right) \\
\Longrightarrow\ &\|\theta^* - \theta_0\| = O_p\left(1/\sqrt{n}\right),
\end{aligned}$$

which proves consistency.

To show asymptotic normality, define $h^+ = \arg\max_{h:g_n(\theta_0+h/\sqrt{n})=\epsilon_n} \Delta'_{n,\theta_0} h - \frac{1}{2} h' H_0 h$. By

Assumption 11,

$$\begin{aligned} h^+ =& \arg\max_{h:g(\theta_0+h/\sqrt{n})=-g_n(\theta_0)+\epsilon_n} \Delta'_{n,\theta_0} h - \frac{1}{2} h' H_0 h \\ =& \arg\max_{h:\bar{G}'h=-\sqrt{n}g_n(\theta_0)+\sqrt{n}\epsilon_n} \Delta'_{n,\theta_0} h - \frac{1}{2} h' H_0 h \end{aligned}$$

for some $\epsilon_n = o_P(n^{-1/2})$. By consistency, $\bar{G} = G_0 + o_P(1)$. Also construct $\bar{R} = R_0$ as in the proof of Theorem 2. Similar calculations as in Amemiya (1985) show that

$$h^+ = \bar{R}\left(\bar{R}' H_0 \bar{R}\right)^{-1} \bar{R}' \Delta_{n,\theta_0} - \left(I - \bar{R}\left(\bar{R}' H_0 \bar{R}\right)^{-1} \bar{R}' H_0\right) \bar{G}\left(\bar{G}'\bar{G}\right)^{-1} \sqrt{n}\left(g_n(\theta_0) + \epsilon_n\right) = (25).$$

Let $\bar{h} = \sqrt{n}\left(\bar{\theta}_S - \theta_0\right)$ and $\bar{G} = G_0 + o_p(1)$. Noting that $G_0' h^+ = -\sqrt{n}\left(g_n(\theta_0) + \epsilon_n\right) + o_P(1)$, we have

$$g\left(\theta_0 + h/\sqrt{n}\right) = \bar{G}'\bar{h} = -\sqrt{n}\left(g_n(\theta_0) + \epsilon_n\right) + o_P(1) = G_0'\bar{h} + o_P\left(\bar{h}\right).$$

Let $\bar{v} = B\left(\bar{h} - h^+\right)$. Following calculations similar to those after (50),

$$\|\bar{v}_1\| = \|G_0'\left(\bar{h} - h^+\right)\| = o_P(1) + o_P\left(\|\bar{h}\|\right) = o_P(1) + o_P\left(\|\bar{v}\|\right) = o_P(1) + o_P\left(\|\bar{v}_1\|\right) + o_P\left(\|\bar{v}_2\|\right),$$

implying that $\|\bar{v}_1\| = o_P(1) + o_P\left(\|\bar{v}_2\|\right)$. Substituting into (50) analogously to (47) and (48),

$$\frac{1}{2}\bar{v}_2'\left(\left(R'R\right)^{-1} R'HR\left(R'R\right)^{-1} + o_P(1)\right)\bar{v}_2 + o_P(1)\|\bar{v}_2\| \leqslant o_P(1)$$

which implies that $\bar{v}_2 = O_P(1)$ and subsequently $\bar{v}_2 = o_P(1)$. ∎

Proof of Theorem 9.

For θ^* denoting a mean value between $\tilde{\theta}$ and θ_0, define $G^* = \frac{\partial}{\partial\theta'} g_n(\theta^*)$. Also define $\Delta_n^0 = \sqrt{n}\tilde{L}\hat{W}_l \ell_n(\theta_0)$ and $H^* = \tilde{L}\hat{W}_l L^{*'}$ for $L^* = \frac{\partial}{\partial\theta'}\ell_n(\theta^*)$. Let $\tilde{D}_n = \left(\tilde{H} + \lambda_n \tilde{G}\tilde{G}'\right)^{-1}\left(H^* + \lambda_n \tilde{G} G^{*'}\right)$. Using the Taylor expansion

$$\tilde{\nabla} + \lambda_n \tilde{G} g_n\left(\tilde{\theta}\right) = \Delta_n^0/\sqrt{n} + \lambda_n \tilde{G} g_n(\theta_0) + \left(H^* + \lambda_n \tilde{G} G^{*'}\right)\left(\tilde{\theta} - \theta_0\right),$$

we can write

$$\sqrt{n}\left(\theta^{-}-\theta_0\right)=\left(I-\tilde{D}_n\right)\sqrt{n}\left(\tilde{\theta}-\theta_0\right)-\left(\tilde{H}+\lambda_n\tilde{G}\tilde{G}'\right)^{-1}\left(\Delta_n^0+\lambda_n\tilde{G}\sqrt{n}g_n\left(\theta_0\right)\right)$$

Note that

$$\tilde{D}_n=\tilde{B}\left(J^{-1}\tilde{B}'\left(\tilde{H}+\lambda_n\tilde{G}\tilde{G}'\right)\tilde{B}\right)^{-1}\left(J^{-1}\tilde{B}'\left(H^*+\lambda_n\tilde{G}G^{*\prime}\right)B^*\right)B^{*-1}=I+o_P\left(1\right)$$

and

$$\begin{aligned}
&\left(\tilde{H}+\lambda_n\tilde{G}\tilde{G}'\right)^{-1}\left(\Delta_n^0+\lambda_n\tilde{G}\sqrt{n}g_n\left(\theta_0\right)\right)\\
=&\tilde{B}\left(J^{-1}\left(\tilde{B}'\tilde{H}\tilde{B}+\operatorname{diag}\left(0_{d_l},\lambda_nI_{d_S}\right)\right)\right)^{-1}J^{-1}\begin{pmatrix}-\tilde{R}'\Delta_n^0\\-\tilde{G}'\Delta_n^0+\lambda_n\sqrt{n}\left(\tilde{G}'\tilde{G}\right)g_n\left(\theta_0\right)\end{pmatrix}\\
=&\tilde{B}\begin{pmatrix}\left(\tilde{R}'H\tilde{R}\right)^{-1}\tilde{R}'\Delta_n^0+\left(\tilde{R}'\tilde{H}\tilde{R}\right)^{-1}\tilde{R}'\tilde{H}\tilde{G}\left(\tilde{G}'\tilde{G}\right)^{-1}\sqrt{n}g_n\left(\theta_0\right)\\-\sqrt{n}\left(\tilde{G}'\tilde{G}\right)g_n\left(\theta_0\right)\end{pmatrix}=(25)+o_P\left(1\right).
\end{aligned}$$

More generally, if we only know that $\tilde{\theta}=\theta_0+O_P\left(n^{-\alpha}\right)$, then $\tilde{H}-H^*=O_P\left(\|\tilde{\theta}-\theta_0\|^{\gamma}\right)=O_P\left(n^{-\alpha\gamma}\right)$ and $\tilde{G}-G^*=O_P\left(n^{-\alpha\gamma}\right)$ for $\gamma\geqslant 0$. Typically for smooth models, $\gamma=1$. Therefore

$$\begin{aligned}
\theta^{-}-\theta_0&=\left(I-\tilde{D}_n\right)\left(\tilde{\theta}-\theta_0\right)+\frac{1}{\sqrt{n}}(25)+o_P\left(\frac{1}{\sqrt{n}}\right)\\
&=\left(\tilde{H}+\lambda_n\tilde{G}\tilde{G}'\right)^{-1}\left(\tilde{H}-H^*+\lambda_n\tilde{G}\left(\tilde{G}'-G^{*\prime}\right)\right)\left(\tilde{\theta}-\theta_0\right)+\frac{1}{\sqrt{n}}(25)+o_P\left(\frac{1}{\sqrt{n}}\right)\\
&=O_P(1)\left(O_P\left(n^{-\alpha\gamma}\right)+O_P(1)O_P\left(n^{-\alpha\gamma}\right)\right)O_P\left(n^{-\alpha}\right)+\frac{1}{\sqrt{n}}(25)+o_P\left(\frac{1}{\sqrt{n}}\right)\\
&=O_P\left(n^{-\alpha(1+\gamma)}\right)+\frac{1}{\sqrt{n}}(25)+o_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

In the κth iteration, then $\theta^{-}-\theta_0=O_P\left(n^{-\alpha\kappa(1+\gamma)}\right)+\frac{1}{\sqrt{n}}(25)+o_P\left(\frac{1}{\sqrt{n}}\right)$. Hence $\sqrt{n}$ consistency can be achieved in at most $\kappa\geqslant 1/\left(2\alpha\left(1+\gamma\right)\right)$ iterations. ■

In appendix C, we consider using numerical derivatives to obtain $\tilde{H}$ and $\tilde{G}$. Numerical

differentiation changes the iteration rate to

$$\theta^- - \theta_0 = \left(O_P\left(\sqrt{\frac{\log n}{n\epsilon_n}}\right) + O\left(\epsilon_n^p\right) + O_P\left(\|\tilde{\theta} - \theta_0\|^\gamma\right)\right) O_P\left(\|\tilde{\theta} - \theta_0\|\right) + \frac{1}{\sqrt{n}}(25) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

The kth iteration will thus improve the rate to, for $\epsilon_n = n^{-\phi}$,

$$\begin{aligned}\theta^- - \theta_0 =& O_P\left(\sqrt{\frac{\log n}{n\epsilon_n}}^k n^{-\alpha}\right) + O\left(\epsilon_n^{kp} n^{-\alpha}\right) + O_P\left(n^{-\alpha k(1+\gamma)}\right) + \frac{1}{\sqrt{n}}(25) + o_P\left(\frac{1}{\sqrt{n}}\right) \\ =& O_P\left(\sqrt{\log n} n^{k(\phi-1)-\alpha}\right) + O\left(n^{-(\phi kp+\alpha)}\right) + O_P\left(n^{-\alpha k(1+\gamma)}\right) + \frac{1}{\sqrt{n}}(25) + o_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Again $\sqrt{n}$ consistency can be achieved in a finite number of steps.

C Numerical differentiation and one-step iteration

In nonsmooth GMM models, numerical derivatives can reduce the convergence rate of $I - \tilde{D}_n$. As in Hong, Mahajan and Nekipelov (HMN) 2015, use step size ϵ_n, appropriate $c_l's$ and unit basis vectors e_j to define

$$\tilde{L} = L_{1,p}^{\epsilon_n} \ell_n\left(\tilde{\theta}\right) = \left\{ \frac{1}{\epsilon_n} \sum_{l=-p}^{p} c_l \ell_n\left(\tilde{\theta} + l\epsilon_n e_j\right), j = 1, \ldots, K \right\},$$

and similarly $\tilde{G} = L_{1,p}^{\epsilon_n} g_n\left(\tilde{\theta}\right)$. The coefficients c_l are determined as $\sum_{l=-p}^{p} c_l l^k = 1$ and $\sum_{l=-p}^{p} c_l l^i = 0$ for $i \neq k$. The following lemma is derived from Theorem 2.37 in Pollard (1984) and Lemma 1 in HMN 2015.

Lemma C.1. *Let $\ell(\theta) = E\ell_n(\theta)$ be more than p times continuously differentiable. Under the conditions in Theorem 2.37 of Pollard (1984), whenever $\log n/(n\epsilon_n) \to \infty$,*

$$\tilde{L} - L_0 = O_P\left(\sqrt{\frac{\log n}{n\epsilon_n}}\right) + O\left(\epsilon_n^p\right) + O_P\left(\|\tilde{\theta} - \theta_0\|^\gamma\right).$$

PROOF OF LEMMA C.1. Decompose $\tilde{L}-L_0=\nabla L_1+\nabla L_2+\nabla L_3$, where $\nabla L_2=L_{1,p}^{\epsilon_n}\ell\left(\tilde{\theta}\right)-\frac{\partial}{\partial\theta}\ell\left(\tilde{\theta}\right)=O\left(\epsilon_n^p\right)$ and $\nabla L_3=\frac{\partial}{\partial\theta}\ell\left(\tilde{\theta}\right)-L_0=\|\tilde{\theta}-\theta_0\|^\gamma$. It remains to show, for $\nabla L_1=L_{1,p}^{\epsilon_n}\left(\ell_n\left(\tilde{\theta}\right)-\ell\left(\tilde{\theta}\right)\right)$,

$$\nabla L_1=\frac{1}{\epsilon_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{l=-p}^{p}c_l\left[\ell\left(X_i,\tilde{\theta}+l\epsilon_n e_j\right)-\ell\left(\tilde{\theta}+l\epsilon_n e_j\right)\right]=O_P\left(\sqrt{\frac{\log n}{n\epsilon_n}}\right)$$

by showing that $\sup_{\theta\in\Theta}\left|\frac{1}{\epsilon_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{l=-p}^{p}c_l\left[\ell\left(X_i,\theta+l\epsilon_n e_j\right)-\ell\left(\theta+l\epsilon_n e_j\right)\right]\right|=O_P\left(\sqrt{\frac{\log n}{n\epsilon_n}}\right)$. Let $\mathcal{F}_n=\left\{\ell\left(X_i,\theta+l\epsilon_n e_j\right),\theta\in\Theta\right\}$. Then for each $f\in\mathcal{F}_n$, $\left(Pf^2\right)^{1/2}\leqslant C\epsilon_n$. For each $M>0$, consider $\varepsilon_n=M\epsilon_n\sqrt{\frac{\log n}{n\epsilon_n}}$. By assumption, $Var\left(P_nf\right)/\varepsilon_n^2\leqslant M/\log n\to 0$, so that for all large n Pollard's symmetrization applies. Furthermore, $\mathcal{F}_n$ being Euclidean, further bound

$$P\left(\sup_{f\in\mathcal{F}_n}|P_n^0f|>2\varepsilon_n\right)\leqslant 2A\varepsilon_n^{-W}\exp\left(-\frac{1}{2}n\varepsilon_n^2/64\epsilon_n\right)+P\left(\sup_{\mathcal{F}_n}f^2>64\epsilon_n\right).$$

Lemma 33 in Pollard (1984) bounds the second term by $C\epsilon_n^W\exp\left(-n\epsilon_n\right)=o\left(1\right)$. Finally bound the first term by

$$AM^{-W}\left(\epsilon_n\sqrt{\frac{\log n}{n\epsilon_n}}\right)^{-W}\exp\left(-\frac{M}{128}\log n\right)\overset{n\to\infty,M\to\infty}{\longrightarrow}0.$$

Therefore, by the symmetrization inequality in equation (30) of Pollard (1984),

$$\lim_{M\to\infty}\limsup_{n\to\infty}P\left(\sup_{f\in\mathcal{F}_n}\frac{|P_nf-Pf|}{\epsilon_n}\geqslant 8M\sqrt{\frac{\log n}{n\epsilon_n}}\right)=0$$

■